EXPLORING THE DIVERSITY IN THE IMPACT OF COLORS OF RATING SCALES ON USER'S RATING BEHAVIOR

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Najia Manjur

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Or

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9
Canada

Abstract

User-generated ratings have become an integral part of data-driven systems, yet they are known to be susceptible to rating bias that can distort the true ratings of users and can subsequently contaminate the effectiveness of the system. Earlier studies have discovered that different characteristics of a rating scale such as granularity, color, neutral point, etc. play a significant role in instigating bias in users' rating behavior. Amongst all, the research works done so far to explore the impact of colors used in rating scales have uncovered significant contradictory patterns of bias in user ratings. This research argues that because of their individuality, users' responses to the influence exerted by the color of the scale are diverse. Personality and culture are known as two consistent representatives of a person's individuality. Yet no attempt has been made to explore the diversity in individuals' responses to the influence of color of rating scales from the perspective of their personality and culture. In addition to it, while investigating the impact of color, the existing research works employed rating scales varying in multiple characteristics other than colors and consequently failed to capture the sole impact of color on users' rating behavior. This research addresses the problem by providing new empirical information about the impact of color-coded rating scales on users' rating behavior. A within-subject study was conducted to collect participants' responses on a demographic and a personality assessment questionnaire and their ratings on different products. The result shows that, extroverts tend to provide biased ratings in star-based scales with contrasting color combinations. On the other hand, collectivists exhibit a tendency to provide biased ratings under the influence of star-based and emoji-based scales with contrasting color combinations. The analysis also revealed some significant directions on how extroverts and collectivists adjust their ratings due to bias. Precisely, taking a personality and culture based approach can help to gain a thorough understanding of the impact of color-coded rating scales on users' rating behavior.

Acknowledgements

I would like to extend my sincere gratitude to my supervisor Prof. Julita Vassileva for the wonderful support and patient guidance she has provided throughout my time as her grad student. Truly, I appreciate her efforts towards the completion of my thesis and consider myself lucky to have such a wonderful supervisor who cares so much about not only the research works but also the mental health of her students. I also acknowledge and thank my advisory committee: Prof. Anh Dinh (external), Prof. Zadia Codabux and Prof. Cody Phillips for their valuable feedback.

I am also thankful to my parents who are my greatest blessings from the Almighty. They always encouraged and inspired me to pursue my master's degree and the sacrifices they made are the reasons I have come so far. I deeply appreciate my husband Minhajul Arifin Badhon for his constant support and encouragement.

Contents

List of Tables

List of Figures

1 Introduction

With the emergence of Web 2.0 (also known as Participative and Social Web), data-driven platforms such as recommender systems, e-commerce websites, online communities, etc., have become the prevalent decision aids in every aspect of our life. They accumulate consumers' post-consumption feedback and leverage that information for providing personalized recommendations tailored to their interests and preferences [47]. User generated feedback tells online vendors about what the consumers care about and offer them scopes for improving services and increasing sales. In reality, recommendations offered by interactive systems significantly impact consumers' decision-making process and shape their online behavior. Statistics show that, over 80% of the contents watched by Netflix subscribers results from the contents recommended by the algorithm[29], YouTube viewers spend 70% of their time watching videos recommended by the system and 30% of the page visited by the buyers on Amazon comes from their recommendations [68].

Post-consumption feedback is usually collected by means of user ratings and reviews. Between them, ratings are apprehended more easily by consumers because they are represented with an uncomplicated numerical format and require a lower cognitive load for a comprehensive understanding of the overall product quality [77]. Explicit consumer ratings are collected utilizing a rating scale, which is a graphical widget with various salient features (e.g. labeling, presentation form, granularity, neutral point, colors, etc.). In the majority of the research works, a common underlying presumption is that consumers' ratings are trustworthy and can be considered as the representative of the true and non-malleable feedback of their experiences with products. It also implies that there is no need to assess the validity of the ratings [5]. However, researchers in behavioral studies have found that, depending on consumers' individual perceptions of different characteristics of rating scales, their decision-making process can be irrational and subject to bias. Consequently, it may cause distortion to their true rating scores. When such biased ratings are accumulated and fed to the system as users' feedback, they can weaken the system's efficiency and compromise the quality of the recommendations suggested by the system. Therefore, understanding the cause and effect of rating bias is of the utmost importance.

Rating scales may also vary in terms of their features and visual appearance across different platforms. For example, YouTube uses a thumb icon-based binary scale (like/dislike), Metacritic uses a circle based scale with numeric labels and Amazon uses a star-based scale to collect consumer feedback. In many spheres, these platforms may require to correctly interpret ratings from other platforms. For example, in order to resolve the cold-start problem and generate recommendations for its new users, a new system may require to import ratings from well-established systems; many hybrid web applications may need to agglomerate ratings

from multiple sources that employ rating scales with dissimilar features. In the aforementioned scenarios, it is important to have an understanding of the diversity in a user's approach to utilizing heterogeneous rating scales for evaluating similar products. Otherwise, the interpreted ratings would not properly reflect individuals' preferences for products. Therefore, in-depth knowledge regarding this matter is required for the system to be able to correctly interpret the ratings instead of implementing a straightforward arithmetic score adjustment technique.

Although many existing research works have addressed the issue of bias by investigating the impact of rating scale characteristics on users' rating behavior [80, 9, 15], none of them addressed the variety in users' responses emerging from their individuality. For instance, earlier research works found that in the presence of a rating scale with a neutral point, some respondents might choose the neutral point only to avoid an extreme response. On the other hand, the absence of a neutral point may compel an actual impartial opinion to choose an extreme score. As a consequence of this influence, users may assign a biased rating score and unknowingly cause distortion to their genuine evaluation [23]. Furthermore, according to [9], the numeric labels of rating scales can also induce bias in user behavior. A scale with labels ranging from +4 to -4 will not be interpreted similarly as a scale ranging from 9 to 1. When the negative points are labeled with negative numbers, they will be perceived as more negative than the negative points labeled with positive numbers. Interestingly, while addressing the impact of colors in rating scales on users' rating behavior, the existing research works found contrasting patterns of rating bias. Tourangeau et al. [75] concluded that, under the influence of scales with endpoints shaded with different colors, respondents would adjust their scores towards the higher end of the scale. On the contrary, Bonaretti et al. [13] hypothesized that, the influence of color would persuade respondents to shift their scores towards milder ratings, or in other words, the central area of the scale.

To date, the previous works drew conclusions from studies designed with a one-size-fits-all approach to investigating bias in user ratings and the contradictory patterns are the consequences of such an approach. These contradictions also accentuate the limitations of not considering the differences in users' fundamental idiosyncratic attributes. Moreover, the color-coded rating scales adopted in the existing works were associated with other scale characteristics (such as numerical or verbal labels) which may have distorted the sole impact of colors in rating scales. Not only for bias detection, earlier research works mostly overlook users' individuality and adopted a generalized mechanism for bias mitigation as well [61, 30]. Whereas adopting a personalized mechanism on the basis of users' individuality can easily narrow down the existing intricate processes.

In summary, there exists a gap in earlier research works on exploring the unique differences in users' responses to the influence of color-coded rating scales. And this work is an important step towards understanding this diversity.

1.1 Objective

The goal of this research is to investigate the diversity in individuals' approaches to utilizing different color-coded rating scales. This will help the system designers to understand whether the one-size-fits-all scales adopted by data-driven systems reflect genuine user-feedback or not, irrespective of the users' idiosyncratic attributes. In the studies of behavioral science, personality and culture are viewed as the stable and consistent representatives of a person's individuality [70, 76]. In view of this, the research in this thesis aims to bridge the existing research gap by taking a personality and culture-based approach to investigate the role of color-coded rating scales in inducing biased rating behavior. The investigation is conducted through a within-subject study which is designed by maintaining the uniformity in different characteristics of the scales, except for color, with an aim to capture the sole impact of color. An analysis is then performed to identify potential rating bias and possible direction of rating score adjustments resulting from the bias.

1.2 Contributions

Consumers' personality and culture have been proved to be effective at helping researchers explore the impact of different rating scale characteristics on consumers' rating behavior. Yet, to the best of my knowledge, no research has attempted to map the role of users' personality and culture to their biased behavior in color-coded rating scales. In addition to conducting an in-depth investigation that was absent in earlier research works, this research makes the following contributions:

1. This research complements and adds to the large body of the existing knowledge about the impact of rating scale characteristics (e.g. granularity, neutral point, presentation form, etc) on individuals' rating behavior in data-driven systems. To be more specific, the user study contributes novel insights into the rating scale utilization styles of users with different personalities and cultures. It provides a more elaborate and specific presentation of results that was missing in the existing research works. To the best of my knowledge, this work is the only contribution made to the research on the sole impact of colors of a rating scale.

2. The results also scrutinize and explain the contrasting rating behaviors of consumers in color-coded rating scales which came to light in the literature reviewed in this thesis, by means of consumers' individuality. As stated in the literature review, researchers in [75] and [13] disagreed on the direction of consumers' score adjustment in similar color-coded scales. The results attribute such behaviors to the personality and culture of consumers and draw attention to the importance of taking users' individuality into account, instead of taking a one-size-fits-all approach for scrutinizing their rating behavior.

3. This research contributes to the rating score conversion mechanism across different platforms. Since strong associations among the rating scales have been explored in the analysis, therefore, ratings on a color-coded scale could be easily predicted, with the provision that users' ratings on an associated scale

are already known. Furthermore, it provides a proof of concept for researchers to avoid a universal, predefined mapping mechanism for interpreting rating scores and to embrace the concept of personalized conversion or translation mechanism.

4. With an aim to eliminating the existing biases and improving the efficiency of the system, this research offers practical guidelines for designers of interactive systems on the grounds of the key findings. It demonstrates the essentiality of taking a more personalized approach to bias-aware system design, instead of considering that a one-size-fits-all approach would be equally effective for everyone.

1.3 Organization of Thesis

The organization of this thesis is as follows:

In Chapter 2, I discuss the background theories and techniques implemented in this research. This is followed by Chapter 3, which presents a review of existing research works relevant to this research and their limitations. Chapter 4 discusses the methodology of the user study carried out to investigate the vulnerability of users with different personalities and cultures to the influence of color-coded rating scales. It also includes the details of the design, development and implementation of the study. Chapter 5 elaborately describes the analysis of data to determine the bias induced from the impact of color-coded rating scales and the possible direction of the rating score adjustment due to the bias. It also presents the discussions on the implications of the results and answers the research questions. Finally, Chapter 6 proposes specific design recommendations and summarizes the limitations and future works of this research.

2 Research Background

The main aim of this research is to contribute to preserving the efficacy of online rating systems by investigating the existence of bias in users' rating behavior across color-coded rating scales and mapping the bias to users' individuality. This entails measuring the statistical difference by comparing the rating scores provided by the respondents using a monochromatic scale against the scores provided using four color-coded rating scales. To accomplish this, I used the Wilcoxon Signed Rank test which is a non-parametric statistical hypothesis test. It provided a concrete result of the statistically significant difference between two related rating scores of each respondent in the absence and presence of bias due to the color-coded scales. However, considering the respondents exhibited a biased rating behavior, the nature of score adjustments because of the bias was yet unforeseeable. Therefore, I further examined the biased rating scores to mine the interesting association rules from them and to acquire a better perception of the rating that is subject to bias and the direction of adjustment for a biased rating score by utilizing the Apriori algorithm. In the following sections of this chapter, I discuss the aforementioned methods used to analyze the data.

2.1 Non-parametric Wilcoxon Signed Rank Test

Non-parametric statistical analysis is an apt method for analyzing a dataset that does not assume a Gaussian distribution. The Wilcoxon Signed Rank test is a non-parametric test for comparing two paired samples to establish whether two populations' mean ranks are statistically significantly different or not. For the result of the test to be correct, the population from which the paired observations are to be analyzed, should not violate one or more of the assumptions. The assumptions are: the measurement scale for the dependent variable is ordinal or continuous, the same subjects are present in both conditions of two related samples to be compared and the distribution of the difference between two related samples is not normal [72]. In inferential statistics, the null hypothesis is the default statement which assumes that the pairs of observations are drawn from the same population, and therefore their means or medians are equal. After performing the significance test, if the null hypothesis is rejected, then it can be suggested that the paired samples were drawn from statistically significantly different populations. The rejection of the null hypothesis refers to the retainment of the alternative hypothesis. When the alternative hypothesis is non-directional and does not specifically suggest which mean is larger, it is called the two-tailed test.

- H_0 (null hypothesis): The median difference is equal to zero.
- H_1 (alternative hypothesis): The median difference is not equal zero.

Let N be the sample size and (X_i, Y_i) be the pairs of observations where $i = 1, 2, \ldots, N$. The steps for conducting the Wilcoxon Signed Rank test are summarized below:

1. Calculate the signed difference $D_i = Y_i - X_i$ and the absolute difference $\mid D_i \mid$ for each paired observation (X_i, Y_i) of the population.

2. Rank each $\mid D_i \mid$ in ascending order such that the smallest score of $\mid D_i \mid$ gets rank 1. When $\mid D_i \mid$ is equal to zero, ignore (X_i, Y_i) and adjust N accordingly. When $\mid D_i \mid$ for two or more paired samples are equal but nonzero, distribute the average of the ranks across the group of T tied ranks and reduce the variance by $\frac{T^3 - T}{48}$.

3. Calculate W^+, the sum of ranks with positive signs, and W^-, the sum of ranks with negative signs.

4. Choose W=min(W^+, W^-).

5. From the table of critical values for the Wilcoxon signed rank test [55], find the probability (two-tailed p-value) of observing a value of W or more extreme with the help of an exact test [67]. Alternatively, if $N > 20$, an approximation to the p-value is calculated based on the normal test statistics Z-score.

$$Z = \frac{W - \mu_W}{\sqrt{\sigma_W - \frac{T^3 - T}{48}}} \tag{2.1}$$

$$\text{where } \sigma_W = \frac{N(N+1)(2N+1)}{24} \tag{2.2}$$

$$\text{and } \mu_W = \frac{N(N+1)}{4} \tag{2.3}$$

The p-value is used to measure how well the sample data support the null hypothesis or reject the alternative hypothesis of the significance test. According to [6], a p-value is the probability of observing an effect as or more extreme than the sample result by random choice, assuming H_0 is true. Let α be the significance level (the probability of rejecting H_0 when H_0 is true). The typical value of α is 0.05. The comparison of p-value with α assists to decide whether the differences between the medians of the paired samples are statistically significantly different or not.

- If p $\leq \alpha$: reject the null hypothesis.
- If p $> \alpha$: retain the null hypothesis.

2.2 Association Rule Mining with Apriori Algorithm

Association Rule Mining is a data mining technique that discovers recurring relationships between itemsets in various data repositories i.e. relational databases, transactional databases, etc. When a set of items occurs frequently, it is called a frequent itemset and a frequent itemset with n number of items are called

frequent n-itemset. The interesting if-then associations that occur between these frequent itemsets are called association rules. These rules help to understand the probability of the occurrence of such associations [66].

The Apriori algorithm is an association rule mining algorithm that mines frequent itemsets from the relational database and then uses the prior knowledge of frequent itemset properties to generate interesting association rules. It employs a level-wise generation of frequent itemsets. To reduce the search space for the level-wise approach and to improve the performance, it uses an important property called the Apriori property, which requires that, all nonempty subsets of a frequent itemset must also be frequent [31]. The algorithm is favorable for small databases [46] and can be utilized in a wide variety of applications. A typical example of such an application can be named as "Market Basket Analysis" where consumers' shopping habits are analyzed to find associations between different products that they frequently purchase together. Table 2.1 depicts a database containing a number of market basket transactions of the consumers of a store.

Table 2.1: Market basket transactions.

Transaction ID	Items
1	{Egg, Apple}
2	{Bread, Butter, Apple}
3	{Milk, Bread, Butter, Chips}
4	{Coffee, Bread, Butter, Sugar}
5	{Coffee, Bread, Butter, Lemon, Honey}

In the transactional database, the itemset {Bread, Butter} is a frequent 2-itemset that often appeared together. It suggests that bread and butter were often purchased together by consumers. The pattern can be presented as an association rule: {Bread}→{Butter} which indicates that a consumer who has bought bread is likely to purchase butter. An association rule has two parts:

- **Antecedent (if):** An antecedent is an item or a group of items found in the data set. In the given example of the rule {Bread}→{Butter}, bread is the antecedent.
- **Consequent (then):** A consequent is the item that occurs with the antecedent. Here, butter is the consequent.

To identify the most interesting associations rules and to evade the possibility of these rules to occur by chance, three different metrics (support, confidence and lift) are used. Rules are considered interesting if they satisfy the predefined constraint of the minimum threshold on support, confidence and lift.

- **Support:** It indicates how frequently an itemset appears in the relational database and identifies the usefulness of the discovered association rules for further analysis. For example, a support of 0.02 for the association rule (2.4) means that in 2% of the total transactions bread and butter were purchased

together by the consumers.

$$\{\text{Bread}\} \rightarrow \{\text{Butter}\}[\text{support}=0.02,\ \text{confidence}=0.6,\ \text{lift}=1.6] \tag{2.4}$$

$$\text{Support}(X \rightarrow Y) = \frac{\text{Frequency of both } X \text{ and } Y}{\text{Total number of entries}} \tag{2.5}$$

- **Confidence:** It is a measurement of the certainty of the association rules and it indicates how likely the consequent will occur, given the antecedent has already occurred. For example, the association rule (2.4) has a confidence value of 0.6. It means that 60% of the consumers who bought bread, also bought butter.

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(X \rightarrow Y)}{\text{Support}(X)} \tag{2.6}$$

- **Lift:** It depicts the correlation between antecedent and consequent of an association rule and indicates the rise in the conditional probability of occurrence of the consequent given the antecedent has already occurred. If lift$<$ 1, the antecedent and the consequent of the rule are negatively correlated which means that the occurrence of one implies the absence of the other. In cases where lift$>$1, the items are positively correlated, which refers that the occurrence of one implies the presence of the other. If lift=1, the occurrences of the items are independent.

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Support}(X \rightarrow Y)}{\text{Support}(X) * \text{Support}(Y)} \tag{2.7}$$

The steps followed in the execution of the Apriori algorithm are [31]:

1. Scan the database to collect the set of frequent 1-itemsets and get the support S for each item. Compare S with the prespecified value of the minimum support threshold S_{min}. Collect the items that satisfy S_{min}. The resulting set of 1-itemsets is denoted by L_1.
2. Generate a set of candidates with k-itemsets, C_k, by implementing $L_{k-1} \bowtie L_{k-1}$, where l_1, l_2, ..., l_i are the itemsets in L_{k-1} and $l_i[j]$ is the jth item in l_i. To ensure an efficient execution, it is assumed that the members of each itemset is sorted in a lexicographical order. The itemset resulting from joining l_1 and l_2 is $\{l_1[1], l_1[2], ..., l_1[k-2], l_1[k-1], l_2[k-1]\}$, given their first $(k-2)$ items are in common.
3. Scan the database to determine the support S for each candidate in C_k. Compare the support S with S_{min} and collect the k-itemsets that satisfy S_{min}. However, depending on the size of C_k, it can be computationally heavy. To prune the size of C_k, apply the Apriori property. According to the Apriori property, if any $(k-1)$ subset of C_k is not in L_{k-1}, then the candidate cannot be a frequent itemset and can be eliminated from the candidate k-itemset. The resulting set of frequent k-itemsets is denoted by L_k.
4. Repeat step 2 to 3 while $C_k \neq \emptyset$.
5. For every frequent itemset l, generate all nonempty subsets of l.

6. For each nonempty subset l_s of l, output the rule $l_s \rightarrow (l - l_s)$, if confidence C and lift L of the rule satisfy the minimum confidence threshold C_{min} and the minimum lift threshold L_{min}. Here, the rule indicates that the occurrence of l_s implies the occurrence of $(l - l_s)$.

3 Literature Review

The accumulation of explicit and implicit feedback data collected from the consumers is crucial for the electronic marketplace to improve consumer satisfaction and system effectiveness [37]. Ratings are the most widely used form of explicit feedback which are collected by means of rating scales and the design of rating scales can manipulate and bias consumers' ratings. Bias in user rating behavior can contaminate the quality of user feedback and consequently, the efficiency of the system can be compromised. Conducting research focusing on the influence of the design elements of a rating scale on the rating behavior of the users, is therefore imperative in order to preserve the integrity and efficiency of the system. To provide a background for my research, I reviewed the literature on the following topics: an overview of user-generated feedback and rating scale characteristics, how the characteristics of a rating scale can impact consumers' rating behavior, the role of personality and culture as the determinants of consumes' rating behavior and the Five Factor Model (FFM) for identifying personality traits.

3.1 Overview of User Generated Feedback

User-generated feedback is the insight provided by the customers on their overall satisfaction and experience with a product or service. It has become an integral and influential part of different domains of the data-driven platforms including e-commerce, social media, employment sites, entertainment and news portals [27]. For instance, a study on Yelp.com found that an increase of one star in a restaurant review can lead up to 9% increase in their revenue [56]. The focus of retail websites and online rating systems is to analyze consumer-generated product evaluation and predict contents related to consumer preferences which can be helpful for the users to make purchasing decisions in the future [4]. To this aim, the mechanism of the system collects explicit and implicit feedback from the users. Explicit feedback is collected in the form of ratings and reviews from the consumers on the products they already had experience with and implicit feedback is acquired through indirect monitoring of consumers' behavior (e.g. number of times a song was played or a movie was watched by a user). They are eventually leveraged to generate consumer profile and improve the predictive accuracy of the algorithm adopted by the system so that the users' experience can be tailored to fit their specific individual requirements. Thus, users' feedbacks help to filter out the unlimited number of alternatives available online and produce personalized recommendations for the users. However, their functionalities are not limited to generating personalized experiences for the users, they also serve as an influential and credible source of information for the consumers and help them assess the risk associated

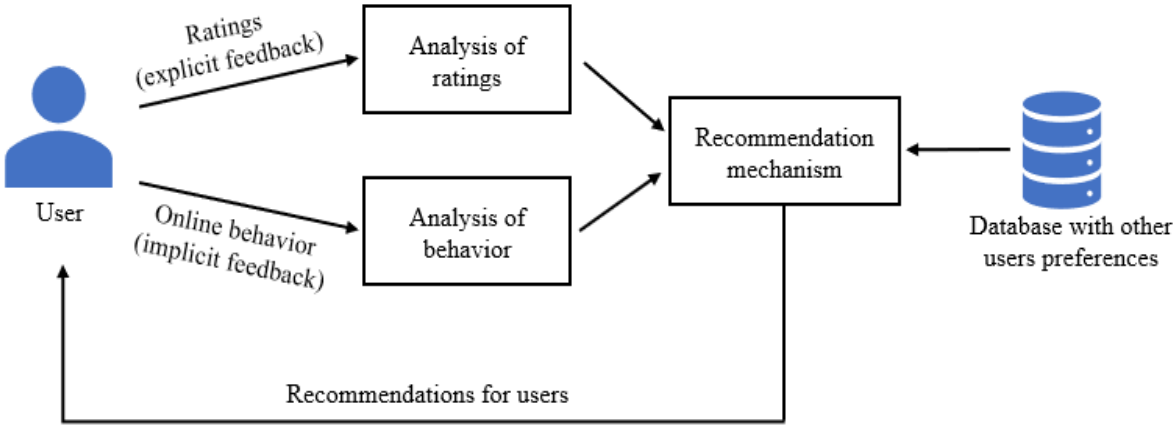


Figure 3.1: Overview of the mechanism of data-driven systems

with making a purchasing decision [50, 12]. Consumer feedback also helps online vendors to advertise their products and services, quantify customer satisfaction, upgrade product quality, make business decisions and create a better customer experience. According to the Nielsen Global Survey conducted on 30,000 consumers in 60 countries in 2015, consumer-generated online feedback is trusted by 66% of customers and ranked as the third most-trusted source of information after recommendations of friends and family and advertisements on branded websites [10]. The mechanism of data-driven systems (such as e-commerce sites, recommender systems etc.) and how they employ explicit and implicit user feedback to help users make a purchasing decision are shown in Figure 3.1.

Explicit feedback provided by consumers can be distinguished in two types of formats: online ratings and reviews. Ratings typically provide product assessment in a numerical format, while reviews in textual format [59]. Because the numerical format is quite straightforward and requires a lower cognitive load for the viewers to gain a comprehensive understanding of the product's quality than reviews, thus consumers mainly prefer to put their trust in online ratings over online reviews especially when they are at the initial stage of the decision-making process [77]. The trust of consumers in online ratings was further emphasized by Gavilan et al. in [24], where a study was conducted with 130 participants who were asked to book a hotel using a website exclusively designed for the experiment. The study revealed that, reviews add values to the credibility of good ratings, but bad ratings are considered credible enough on its own regardless of the number of reviews. The trustworthiness of online rating systems lies in the quality of ratings they deliver. Therefore, identifying the factors degrading the quality of the ratings is critical to assist the system in maintaining its quality and generating efficient and relevant recommendations for the users.

Ratings provided by the customers are expected to reflect their individuality and represent their truthful and diversified opinions on their perceived product quality. However, a human's decision-making process is not always rational. Due to the cognitive limitations of the human mind, this process can be subject

to bias evoked by the system mechanism or by the issues present in the design elements adopted by the system [4, 5, 18]. As a result, customers may deviate from their true feedbacks and produce biased online ratings [77]. According to Sackett, the word "bias" is defined as, "any process at any stage of inference which tends to produce results or conclusions that differ systematically from the truth" [74]. For example, one of the highly discussed forms of bias is the "Anchoring Effect" theorized by Tversky and Kahneman where an individual displays a tendency to rely on the initial piece of available information and produces decisions that are biased towards that value [78, 62]. Examples of numerous studies are available confirming the existence of different user bias including the anchoring effect. For instance, in an experiment conducted by Zhang et al. [84], participants were asked to watch an episode of a TV show and provide their ratings for it. During the rating collection process, when an artificially generated high rating score was recommended with the rating scale, it manipulated participants' judgment and provoked them to provide higher ratings even when their actual experience was not worthy of it. The effect appeared to be more prominent among viewers who received the artificial rating prior to their consumption than viewers who received it at the post-consumption stage. In this way, tempered and non-rational feedback from the consumers reduced the credibility of the recommendation system.

My research exclusively focuses on exploring the influence that the design elements (also known as the characteristics) of the rating scales can exert on consumers and their possible behavioral outcome due to that influence in the context of online rating systems.

3.2 Categorization of Rating Scale Characteristics

A rating scale is a graphical widget that is used to facilitate the consumers to provide their ratings as a form of explicit feedback [15]. Rating scales are widely used across numerous numbers of websites and their characteristics may differ from each other in various factors including the choice of visual representation, number of score intervals, type of labels used to represent the interval points, etc. For example, Netflix uses a binary rating scale which exploits a visual metaphor of thumbs up and thumbs down icons, Tripadvisor uses circles, LateRooms uses square-shaped icons and eBay uses a 5 star-based scale to collect user feedback. Both Amazon and Trustpilot are similar to each other in terms of the visual metaphor and the number of scale interval but they still differ in their choice of color. Amazon uses a star-based monochromatic scale with yellow color, whereas Trustpilot uses a star-based scale shaded with the combination of red-yellow-green colors. Variations are also evident in the design norms of rating scales adopted by websites representing the same industry. For example, in the context of movie recommender systems, IMDb uses a 10-point star-based rating scale, Filmeter uses a rating scale of 6 points with expressive icons representing a hypothetical viewer [15] and Rotten Tomatoes uses a 5 star-based rating scale.

According to Van Barneveld et al. [79], the main features for designing an interface for presenting system-generated prediction and collecting user-generated feedback can be categorized into four groups:

- **Presentation form:** A rating scale can use different visual concepts to represent itself. For example, a scale may adopt a numerical score format or a group of symbols with visual metaphors (e.g. emoji) to facilitate the rating activity.
- **Scale of prediction:** A rating scale can adopt continuous or discrete measurement (based on the choice of numbers to represent the scale data), can vary in range (based on the granularities or the number of available options e.g. 1 to 5 or 1 to 10), can vary in precision (e.g. 1,2,3 or 1,1.5,2,2.5,3), and can be symmetric (having endpoints in both positive and negative values, e.g. -3 to 3) or asymmetric (represented with only positive numbers e.g. 1 to 5).
- **Visual symmetry or asymmetry:** The visual representation of positive and negative endpoints of a rating scale can be symmetric or asymmetric. For example, a symmetric scale (e.g. -2 to 2) can be visually asymmetric if it is represented with 5-points emoji scales with a different emoji representing each interval and the third emoji referring to the central position of the scale.
- **Color:** A rating scale may use different colors to make the endpoints appear visually distinguishable (e.g. using a color combination of red-yellow-green to represent the low-neutral-high scores of a scale).

Based on the abovementioned characteristics defined in [79], rating scales are classified into three main types by Cena et al. [16] through visual inspection. Table 3.1 shows a brief comparison among three clusters of rating scales concerning their characteristics. The clusters are described below:

Table 3.1: Comparison among three clusters of rating scales defined in [16].

	Clusters		
Rating scale characteristics	Human scales	Neutral scales	Technical scales
Visual metaphor	Smileys,thumbs	Stars	None
Labels	No	No	Numerical
Granularity	Low	High/low	High
Neutral point	No	Yes	Yes
Negative points	Yes	No	Yes
Emotional connotation	Yes	No	No

- **Human Scales:** Scales in this category are represented with human visual metaphors (e.g. smileys, thumbs) and convey strong visual characters. They influence users to make judgments based on the emotional connotations of the icons used in the scales rather than making any precise quantification. Every point of the scale is usually depicted with a different icon. Human scales usually do not use labels and explicit neutral points, have low granularity and use negative points.
- **Neutral Scales:** Scales in this cluster are considered as a widely used standard form of rating scales. They do not use any visual metaphors and therefore depict no strong emotional connotation. Unlike

"Human Scales", every point of a neutral scale is represented with the same icon (e.g. stars, circles). They usually have neutral points, can adopt both low and high granularity but do not use labels and negative points.

- **Technical Scales:** They mainly focus on quantitative evaluations of items. When necessary, scales in this cluster might adopt technological metaphors which do not convey any emotional connotation (e.g. Likert scales, sliders). Technical scales usually use numerical labels, neutral point and negative points and have higher granularity than the rest.

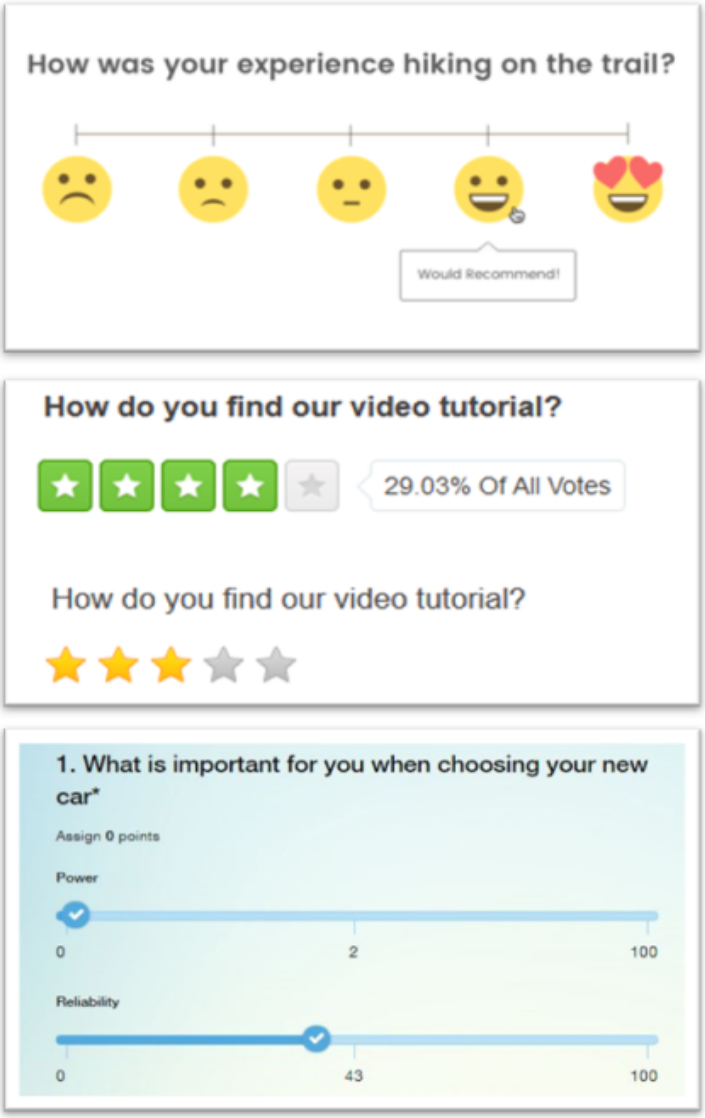


Figure 3.2: Examples of Human Scale, Neutral Scale and Technical Scale (from top to bottom)

Figure 3.2 depicts examples of rating scales from each of the three clusters defined in [16]. The combination of the features mentioned in [79] forms the "personality" of any rating scale. The "personality" of a rating

scale is defined as "the way rating scales are perceived by users and affect their behavior" in [26]. In order to design the rating scales employed in my study, I utilized the set of features on the clusters mentioned in this section.

3.3 Impact of Rating Scale Characteristics on Rating Behavior

Several behavioral studies suggested that subject's judgments can be influenced by the elements of the experimental environment [5]. This implies that different characteristics of a rating scale can manipulate users' opinions and hence variations in users' feedback for the same product or service can be observed across different platforms. Let us consider the case of a movie's rating in the context of different movie rating systems: The movie Captain Marvel (2019) received 6.9 out of 10 in IMDb which uses a 10-point star scale and 3.5 out of 10 in Metacritic which uses a 10-point circular rating scale that transitions into red-yellow-green color triplets upon hovering. Despite having similar granularities, the ratings for Captain Marvel in IMDb is higher than Metacritic. Given that the movie they rated is a constant, such variance in the overall evaluations may be subject to the variations of colors used in the rating scales adopted by IMDb and Metacritic. This draws attention to the influence that the design elements of a rating scale may exert on the user's rating behavior. For example, Adomavicius et al. mentioned in [4] that users' rating behavior is susceptible to bias depending on the design of the rating interface employed to collect their explicit feedback. In their study, 287 participants were asked to rate 50 jokes using different rating scales which displayed different system generated ratings. It was observed that the participants adjusted their ratings in accordance with the appearance of the rating scale. The authors concluded from their study that neutral scales associated with a numerical form of the score is more capable of inducing bias in user behavior than the slider or binary (thumbs up/ thumbs down) rating displays. Similar studies have also been conducted in [5, 18] to investigate the same issue. In what follows, I will address the existing research works accentuating the bias that can be induced by the design characteristics of a rating scale.

3.3.1 Impact of Scale of Prediction on Users' Rating Behavior

Although the impact of the number of available points in the scale (granularity) on user behavior is a widely discussed topic, research works have exhibited inconsistent results in this context. The researchers in [53] asked 137 participants to assess their overall happiness using scales varying in granularities (4, 5,7 and 11-point Likert scales). By considering the normalized values using an 11-point Likert scale as a baseline, the authors compared the value of mean ratings across different rating scales. The mean rating in the 11-point Likert scale was observed to be higher compared to the rest. It indicates that a rating scale with a higher granularity will produce higher ratings than a scale with lower granularity assuming that they are indifferentiable in other features. In another experiment conducted in [26], participants were asked to rate five food courses they liked using a rating scale of their choice. The rating scales employed in the study were

designed by adopting different variations of four different features: granularity, range, numerical labels and the presence of an intermediate position. In contrast to [53], it was observed that the use of a 3-points star-based scale encourages the user to rate higher compared to a 5-star rating scale. Kuflik et al. [49] observed a similar result in an experiment conducted at an archeological museum where the users rated various presentations using five randomly presented rating interface exclusively designed for the experiment. The authors concluded from the study that a rating scale with a coarser granularity manipulates the raters to avoid the extremely negative scores and as a result, they tend to rate higher than the average. However, the behavioral pattern observed in [53] contradicts the pattern in [26, 49].

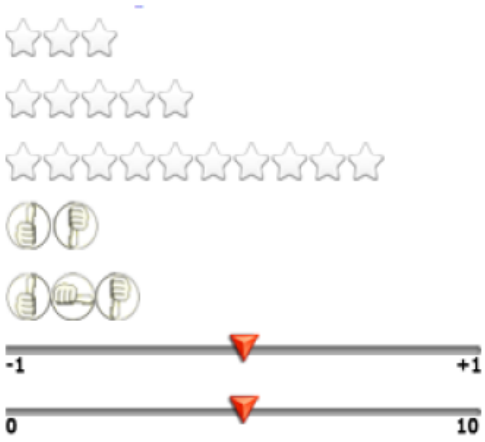


Figure 3.3: The rating scales adopted in the user study of [26]

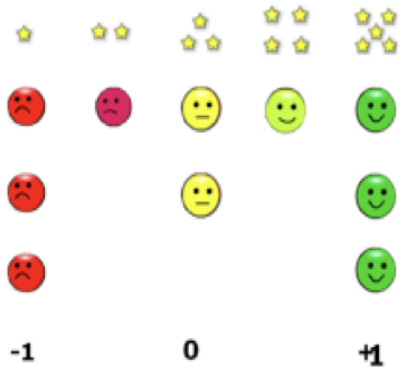


Figure 3.4: The rating scales adopted in the user study of [49]

According to [26, 9], the explicit presence of negative numerical labels on a rating scale also exerts an influence on consumers' rating behavior. In a rating scale labeled with both positive and negative numerical points, the positive side of the scale acts as a cognitive anchor for the raters and is perceived as a safer

alternative compared to the negative side. Consequently, users rate higher than they would in an unlabeled rating scale. In a different study conducted in [81], an overall impact with a similar direction on consumers' score adjustment was identified in rating scales with different labeling formats. The scale where all points were labeled produced a more moderate response compared to the scale where only the endpoints were labeled.

A study conducted in [23] has shown that the presence or absence of a neutral point in a decision-making interface produces distorted ratings. Rating scales with a neutral point may induce bias and force the users with negative opinions to choose the neutral position instead. The bias is caused by users' inclination to choose less extreme opinions over extremely negative opinions. On the other hand, the absence of a neutral position in rating scales exerts a contrasting influence and manipulates users with a neutral opinion to adjust their rating to the upper endpoints of the scales. In such a context, users are prone to give higher ratings.

3.3.2 Impact of Presentation Form on Users' Rating Behavior

Figure 3.5: The rating scales adopted in the user study of [15]

Examples of research exhibiting the impact of "human" scales on users' rating behavior are few and far between. Since the "human" scales use the smiley face as a visual metaphor, they have a strong emotional connotation. Cena et al. [15] conducted a user study about the indoor navigation of an archeological museum with the visitors. In the experiment, the visitors had to rate a multimedia presentation of their choice from a number of available options using emoji-based rating scales with different granularity (2, 3 and 5 points) and a 5 star-based rating scale. Although users' rating behaviors in the star-based scale and 5-points "human"

scale corresponded with each other, the ratings of the "human" scales with different granularity did not. An emoji-based scale with coarser granularity produced a higher average rating compared to the 5-points emoji scale, possibly because the neutral point of the first scale acted as a cognitive cue for the participants to avoid the extremely negative values with sad emoji. To observe the within-industry rating patterns, the authors collected movie ratings from different movie recommender systems. The analysis showed that Filmeter which uses a 6-points based human scale to collect user feedback exhibited a lower average score than other emoji-based scales with finer granularity. In this case, the icons of Filmeter provoked the users to adjust their ratings towards the lower endpoint. According to the user study conducted in [49], "human" scales produced higher average scores compared to the average scores in other scales adopted in the study.

3.3.3 Impact of Color on Users' Rating Behavior

Another important feature of a response scale is color that requires attention in the context of bias in rating behavior. Different websites such as Metacritic, Instantgo, Trustpilot, etc. facilitate their users to rate using color-coded rating scales because the inclusion of color in a rating scale makes the scale more self-explanatory and the endpoints more distinguishable [79]. Surveys conducted in [79] and [57] have found the participants of their studies to be receptive to color-coded scales despite the difference in colors used on them. The survey in [79] referred to users' preference for a combination of warm versus cold colors (e.g. a traffic light model or red-yellow-green trio). On the other hand, the authors in [57] observed users' acceptance for a color-coded star rating scale with red color representing the negative scores and blue color representing the positive scores. The acceptance rate of this scale was higher among the participants compared to other scales (e.g. a binary scale with visual metaphor, a 10-point scale and a 100-point scale) because of its simplicity and easily distinguishable endpoints. Overall, it was suggested to use a warm versus cold color combination to refer to the negative and positive endpoints of a color-coded scale.

Despite the acceptance of color-coded scales among raters, experts have suggested not to use color on a response scale because it can bias their responses [39, 14]. Very few studies have investigated the influence of color on users' interpretation of a rating scale. Tourangeau et al. [75] conducted a web survey where participants recorded their responses to a questionnaire consisted of 16 questions using a survey response scale. The negative and the positive endpoints of the response scale were shaded with different hues of a warm and a cool color respectively (e.g. red and blue). The overall response pattern of this scale was higher than an average rating scale shaded with different hues of a single color. It was observed that associating different hues of two different colors with the response scale confused the users' perception of the subjective distance of the endpoints of the scale. As a consequence, the subjective distance of the scale was perceived as longer than usual by the users and they adjusted their scores towards the higher end of the scale.

A study was conducted in [13] to collect users' evaluation of a lodging experience. In contrast to [75], it hypothesized that the score adjustment would be directed towards the central area of the scale. According to the authors, depending on the context, different colors used in scales can act as cognitive factors that

can influence the numerical interpretation of the scale. Based on the perceived interpretation, users adjust their rating scores on that scale. Thus, users' rating scores can get biased. For example, the presence of green color at the positive endpoint of a scale enhances the positive emotional valence and vice versa on the negative endpoint of the scale. Hence they influence the numerical interpretation of the scale. However, according to the authors, this influence is likely to direct the score adjustment towards milder ratings, or in other words, the central area of the scale. Similar to the research works investigating the effect of rating scales' granularity, the conclusions reached in the aforementioned studies are contrasting.

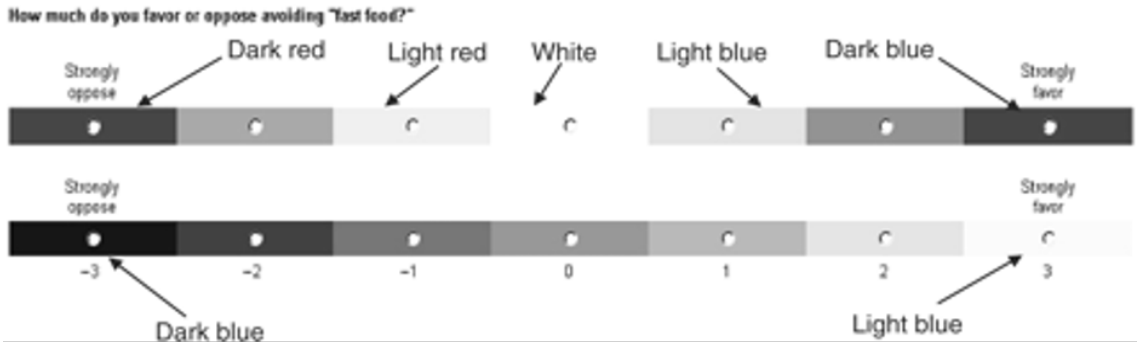


Figure 3.6: Two experimental scales adopted in the user study of [75]

Figure 3.7: The six treatments adopted in the user study of [13]

As summarized in Table 3.2, the literature reviewed so far evidently uncovered some contrasting patterns of bias in users' overall rating behavior by focusing on the general population. Because of their individuality, users respond differently to the design metaphors embedded in a rating scale. Hence, not many uniform patterns were observed in the reviewed literature, which only emphasizes on the essentiality of investigating users' rating behavior at an individual level. Also, there exists a research gap on the sole impact of colors of rating scales since the rating scales adopted in the user studies of earlier research works ([75] and [13]) were associated with different characteristics such as numeric and verbal labels other than colors. Personality and culture are two very stable and consistent aspects of users' individuality [70, 76]. Therefore, with an aim to

bridge the existing gap, this research investigates the sole impact of colors of rating scales on users' rating behavior by taking a personality and culture-based approach.

Table 3.2: Brief overview of the contradictory patterns observed in the literature.

Rating scale characteristics	Source	Result	Comparison
Granularity	Gena et al.[26]	Scales with coarser granularity (e.g.thumbs up/down) tend to produce higher ratings than scales with finer granularity.	Behavioral pattern in [53] contradicts the patterns in [26, 49].
	Kuflik et al. [49]	Scales with coarser granularity tend to produce higher ratings than scales with finer granularity.	
	Lim et al.[53]	Scales with finer granularity(e.g. 11-point Likert scale) tend to produce higher ratings than scales with coarser granularity.	
Color	Tourangeau et al.[75]	Color coded rating scales influence consumers to adjust their rating scores towards the higher end of the scale.	The directions of consumers' score adjustment for similar scales contradict with each other.
	Bonaretti et al. [13]	Color-coded rating scales influence consumers to adjust the score towards the central position of the scale.	

3.4 Five Factor Model of Personality

Prior to establishing the relationship between consumers' personality and their rating behavior, their personality traits need to be identified first. According to the American Psychological Association, "Personality refers to individual differences in characteristic patterns of thinking, feeling and behaving" [8]. To identify an individual's personality type, I have decided to employ the Five Factor Model in this study. It is currently one of the most comprehensive and widely employed models in psychology and its capability to predict personality types and human behavior across different populations is well-established [41]. The Five Factor Model of personality (FFM), also known as the Big Five Model or the OCEAN model, is a hierarchical

formation of the personality traits that serves as the building blocks of personality. It categorizes personality into five broad dimensional traits: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism [58]. An individual's personality is a combination of the five traits. For example, in a personality acquisition test, an individual may score as openness to experience 45%, conscientiousness 55%, extraversion 80%, agreeableness 60% and neuroticism 40% . In this case, extraversion is recognized as the dominant trait while the individual may also exhibit other traits as well [43]. The characteristics representing each of the five traits are depicted in Figure 3.8 and briefly described below:

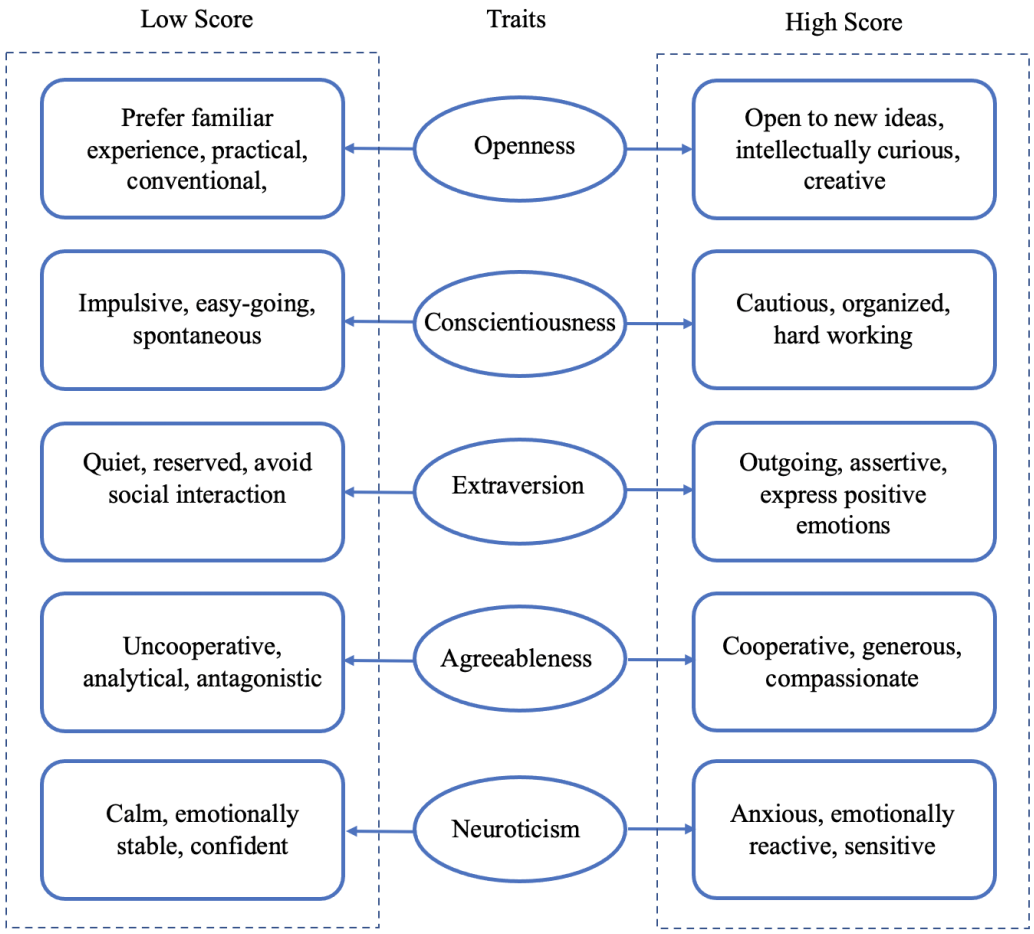


Figure 3.8: Five Factor Model of Personality

- **Openness to Experience** represents an individual's appreciation for new ideas and experiences. People with high openness to experience are open to new ideas, creative and intellectually curious. Conversely, people with high closeness to experience are practical, prefer conventional and familiar experiences and have narrow range of interests.
- **Conscientiousness** represents how individuals control and regulate their impulses. People with high

conscientiousness tend to be cautious, hard-working and exhibit organized behavior. People with lack of direction, on the other hand, tend to be impulsive, disorganized and spontaneous.

- **Extraversion** refers to an individual's fondness for social engagement. Extrovert individuals tend to be assertive, express positive emotions and enjoy interacting with others while introverts tend to be quiet, reserved and avoid social interaction.
- **Agreeableness** is associated with an individuals' tendency to maintain social harmony with others. An individual with high agreeableness is cooperative and exhibit compassionate behavior whereas an antagonistic individual is uncooperative, analytical and exhibit antagonistic behavior.
- **Neuroticism** represents individuals who are prone to negative emotions. Neurotics are unable to deal with stress. They are emotionally reactive and their bad mood strongly impacts their decision-making behavior. On the other hand, people with low neuroticism are calm, emotionally stable, even-tempered and exhibit confidence in their behavior.

To categorize participants into the five personality domains, this research adopts the Big Five Inventory (BFI-44) which is a personality assessment test consists of 44 questions that measure an individual on the basis of the Big Five Factors of personality [41]. In this questionnaire-based test, a participant is asked to report about himself by answering questions such as "Is reserved?", "Is easily distracted?", etc. using a 5-point Likert scale with textual labels starting from "Disagree strongly" to "Agree strongly".

3.5 Personality as a Determinant of Rating Behavior

Earlier research works in behavioral science stated that, personality traits can be held responsible for the variability in preferences and decision-making behavior of individuals [21]. Personality has therefore been used to successfully predict the underlying patterns in user behavior in a wide variety of domains including social networks (e.g. Facebook) [64, 11, 48], microblogs (e.g. Twitter) [63], emails [69], mobile phone usage [17], videogames [35] etc.

Many studies have performed between-group comparisons by taking users' personality type as a factor and provided substantial evidence regarding the link between users' personality and their behavior in online rating systems. For example, in [28], Golbeck et al. retrieved the viewing and rating history of 73 Netflix users and observed that users with high conscientiousness trait are more inclined to give more positive ratings than others, specifically to movies that were previously recommended to them. Researchers noticed a tendency of giving higher ratings among users with high agreeableness [44, 35, 45]. In a survey conducted in [44], the authors mapped users' personalities to their recorded ratings from 17 different genres of movies in the MovieLens data set and implemented a category-wise comparison of ratings with regard to different personality traits. They found that users with high agreeableness exhibited favoritism towards movies of some specific genres (e.g. high rated popular children movies) and rated them slightly higher compared to

users with low agreeableness. Hu et al. [35] recruited 122 consumers and asked them to rate at least 30 items using a binary rating scale (e.g. like/dislike) on a web-based platform that utilized the data set of gifts.com. According to their findings, consumers high in agreeableness tend to give more positive ratings and consumers high in openness tend to rate more items than others. Another study conducted by Karumur et al. [45] mentioned findings that are in line with the results obtained from their work in [44]. The authors extracted the personality profiles and activity logs of 1008 MovieLens users. The activity logs included information about their movie ratings from different genres for four months. The findings of the study revealed that users with high agreeableness were typically inclined to give more positive ratings on average compared to users with low agreeableness trait. User personality evidently plays such a significant role in influencing consumers' online rating behavior that researchers were successful in differentiating between good and bad reviewers on account of the relation between user personality and their review behavior in online communities. For example, Adaji et al. [3] studied the personality traits of consumers who posted unhelpful reviews on Yelp.com and found that individuals providing unhelpful reviews were mostly personalities with high neuroticism. Thus, the authors were successful to identify the bad reviewers on the basis of their personality traits.

Although the above-mentioned studies validated the importance of users' personality traits in differentiating their rating behavior, neither of them has taken design cues of a rating scale into consideration to investigate personality-wise rating behavior. The only research to explain the contrasting rating patterns in scales with different granularities from the perspective of personality traits was conducted by Karumur et al. in [45]. The authors observed that, when given the opportunity to use a rating scale that allowed providing half-star ratings, consumers with high conscientiousness showed an inclination to rate lower than they would, using a regular star rating scale. Interestingly, using the same scale, extroverts showed a tendency to rate higher than the usual. As an explanation for this behavior the authors mentioned that because of being overly cautious, users with high conscientiousness trait provided ratings that were rounded to the nearest point below the ratings they might actually give using a standard star-based rating scale. On the other hand, extroverts, due to their assertive and enthusiastic nature, rounded up their original half-ratings for the movies they liked to the next available higher point in the scale. The findings of the study made it critical to investigate if the users' rating is a personality-driven outcome of their tendency to rate in a particular way under the influence of a particular design metaphor embedded in the rating scale.

Personality traits are also correlated with their preference for colors. According to [82], individuals with low agreeableness exhibit a preference for dark blue, users with high emotional stability have a preference for dark green color and blue is important for both extroverts and introverts. According to [2], every color individually carries an emotional property. But because of being preferred by an individual, the color may enhance the positive valence and thus will most likely play the role of a cognitive enhancer for that user. On the other hand, the absence of a preferred color could result in the absence of the positive valence. Thus, the interpretation of the scale may vary among individuals based on their color preferences. However, users'

susceptibility to the effect of such a cognitive impact may vary according to their personality traits. For example, the magnitude of the color effect could be higher for raters with high openness and extraversion because they are more receptive to new experience than others, on the other hand, it could be lower for raters with high conscientiousness and emotional stability because they are more cautious and even-tempered than others. Thus, because of the variance in different personality groups' interpretation of color, the general approach taken by the existing works might not work. Therefore, this research suggests that scrutinizing users' rating behavior based on their personality types can uncover the underlying patterns of their biased ratings.

3.6 Culture as a Determinant of Rating Behavior

In the cross-cultural context, Hofstede's framework is one of the most widely used models which considers five dimensions of cultural values: individualism-collectivism, power distance, uncertainty avoidance, masculinity-femininity and long-term orientation [32]. Among them, my research highlights the individualism-collectivism aspect since it best captures the variance in the global population [51]. Collectivists tend to make decisions driven by the goal of community benefit whereas individualists make their decisions focusing more on their personal goal [34]. Hofstede's individualism score refers to the degree to which a society values the ties between an individual and the society [1]. A country with a high individualism score will show an inclination towards the individualistic culture, on the other hand, a country with a low score in this dimension will show an inclination towards the collectivistic culture. According to Hofstede et al. [33], most of the north-western countries (e.g. North America, Northern Europe) are individualistic, while Asia, Africa, the Middle East, Mediterranean Europe and Latin America are collectivists.

Although I previously discussed that culture is an important aspect of users' individuality, very few studies investigated the impact of individuals' cultural background on their rating behavior. For instance, in a survey conducted by Lindgaard et al. [54], 40 Canadian (individualists) and 40 Chinese (collectivists) participants assessed a number of homepages on the basis of their visual appeal. The survey revealed that Canadians were inclined to use the lower end of the scale more often than Chinese participants, while the Chinese participants were more inclined to use the higher end of the scale. It indicated that collectivists tend to rate higher than individualists. According to Jenkins et al. [38], collectivists tend to provide moderate ratings more often compared to individualists. Cross-cultural researchers in human-computer interaction (HCI) have found considerable differences between the consumer behavior of individualists and collectivists in online review systems. However, examples of insightful literature explaining bias in cross-cultural rating behavior in light of the designs of rating scales are somewhat sparse. In terms of design variations, collectivists and individualists have very different choices because culture affects individuals' aesthetic preferences. Researchers have shown that [60, 73], people from collectivist culture prefer interfaces which are colorful and visually appealing to the ones that use gray, whereas individualists do not exhibit any such preference. According to Sun et al.

[73], this is because individualists are mostly bothered about information organization than the colors of the interfaces. This research argues that these preferences may convey a positive emotional valence which in turn can play the role of a cognitive enhancer. As a consequence, the numerical interpretation of the scale will get influenced and biased rating decisions will be made. Since the interpretation of colors varies in individuals with different cultures, therefore it can be said that the behavior observed in the existing works regarding the impact of color might not be applicable for everyone. Hence, it is important to investigate the impact of design metaphors embedded in the rating scale on the cross-cultural consumer response.

Investigation of user behavior at an individual level was disregarded in the contemporary studies discussed in the related works despite the prevalent contradictory patterns in consumers' rating behavior in the context of color-coded rating scales. Furthermore, the rating scales adopted in the earlier research works varied in multiple characteristics that might cause distortion to the sole impact of colors of a rating scale. To bridge this gap, this research aims to address the issue of consumers' behavioral bias in color-coded rating scales in more depth by designing and conducting an experiment that takes a personality and culture-based approach.

4 Research Methodology

User-generated product ratings act as a means of advertisement for online vendors to help them promote the overall quality of the products and as a source of information for consumers by providing the assessments of the purchased products given by other consumers. The impact of online ratings is of such great significance that Albornoz et al. mentioned in [20] that, without any hesitance, consumers are willing to pay about 20% more for a service with the highest rating score than for a similar service with a slightly lower rating. If there is bias in the post-consumption ratings of consumers, online rating systems would fail to deliver trustworthy and efficient information regarding product evaluation. This research aims to provide empirical evidence on whether the ratings generated by consumers are the reflection of their genuine experience or the consequence of their biased behavioral pattern by conducting a user study. In this chapter, I describe the methodology including the design rationale and the stages of implementation of the user survey which was conducted to address this issue.

4.1 Research Questions

Online platforms leverage rating scales to collect consumers' post-consumption feedback. As I discussed in Chapter 3 that distinct characteristics of rating scales can impact users' rating behavior. However, the biased behaviors discussed in many of these existing works differed in many respects from each other because they generally took a one-size-fits-all approach while investigating the impact of rating scale characteristics on consumers' rating behavior. For example, [75] mentioned that, the score adjustment in case of a biased rating in a color-coded rating scale, would be directed towards the higher end of the scale, while according to [13], the score would be adjusted towards the central area of the scale. The dissimilarities found in user behavior accentuate the incompetence of the one-size-fits-all approach and the importance of an individuality based approach to investigate bias in color-coded rating scales.

Although personality and culture were used as the behavioral determinants to explain consumers' overall rating behavior in many aspects, no such attempts had been taken by the researchers in the context of rating bias in color-coded rating scales. Hence, it yet remains unclear whether users' uniqueness in terms of their personality and culture would manipulate them to utilize different color-coded rating scales differently for the assessment of the same product. Furthermore, there exists a gap in the related works on exploring the sole impact of colors in a rating scale which needs to be addressed as well.

In this work, I attempted to fill this gap and provide insights on whether, and how the colors embedded

in a rating scale can exert influence on consumers' rating behavior at an individual level. This research has presented reasons to believe that depending on the diversity in the interpretation of color by individuals with different personalities and cultures, the impact of color-coded rating scales will vary. According to the literature reviewed so far, based on the raters' personality based color preferences, cross-cultural color preferences and the variance in the receptivity to the influence according to different personality types, individuals may or may not be susceptible to the cognitive impact of color. With the help of this research, I aim to answer the following questions:

1. Do consumers with different personality traits utilize similar color-coded rating scale differently for the same product?
2. Do collectivist consumers utilize similar color-coded rating scale differently from individualist consumers?
3. In case of a biased rating, how do consumers adjust their actual ratings?
4. Can a personality and culture-based approach clarify the contradictory rating behaviors observed in the literature review?

To answer the aforementioned questions, I designed and implemented a user study in order to collect consumer ratings. In the following sections, I describe the steps taken to design, develop and conduct the user study. This research received a behavioral ethics approval from the Research Ethics Board of the University of Saskatchewan with the approval number BEH 1521.

4.2 Research Framework

This exploratory research aims to determine how a color-coded rating scale can impact consumers' rating behavior at an individual level. A user study was designed by integrating a set of questionnaire and rating activities which would assist the decision-making process regarding the bias induced in consumers' rating behavior by rating scale characteristics. In order to make the user study more available and accessible, it was deployed on the web. Figure 4.1 presents the diagrammatic summary of the design, implementation and analysis of the user study. The research process comprises of three major steps:

1. **Selection of rating scales:** In the first stage, taking the insights from the literature review into account, I decided the rationale of the design of the rating scales. The rating scales selected for this study are the representatives of the scale clusters defined in [16] and were designed based on the feature categorization described in [79]. The experimental setup consisted of five different rating scales (a baseline scale and four experimentally designed scales) to facilitate the participants of the user study to provide their post-consumption ratings for different products. The detailed rationale for the selection of the rating scales is mentioned in Section 4.3.

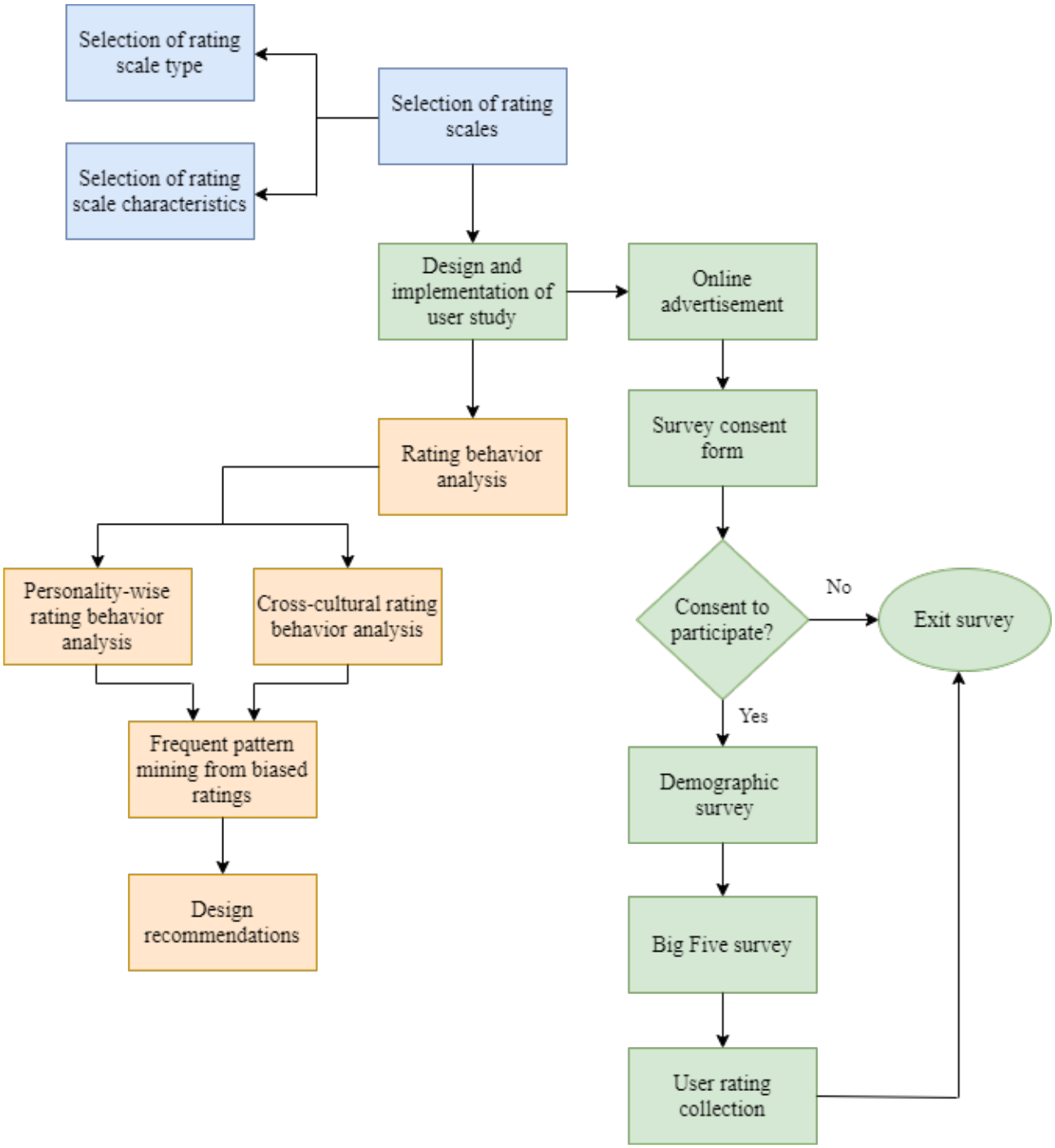


Figure 4.1: Summary of design, implementation and analysis of the user study

2. **Design and implementation of user study:** This step involves the implementation and development of the user study as a web application to make it more accessible to the targeted audience. To complete the study, a respondent had to participate in three different activities: firstly, respondents completed a demographic survey where they answered a number of questions regarding their demographics which

would help to classify the individuals into collectivist and individualist culture. In the second step, participants would answer a set of questionnaires from the Big Five Inventory (BFI-44 items) which would generate their personality profile. In the last stage, participants would provide their post-consumption ratings for different products using the rating scales (a baseline scale and four experimental scales) designed in the first stage. The stages of implementation and the strategies adopted to encourage user participation are described in detail in Section 4.4.

3. **Rating behavior analysis:** In this stage, I present a fulsome analysis of the data collected from the user study and the necessary decision-making process. The decision making process consists of three sub-stages: firstly, I aim to answer research question 1 from the analysis of personality-wise rating behavior and then, research question 2 from the analysis of cross-cultural rating behavior by using the non-parametric Wilcoxon Signed Rank test. These analyses jointly contribute to making a conclusive decision regarding whether consumers' personality and culture would manipulate them to provide biased ratings in a color-coded rating scale or not. Nevertheless, they do not provide a clear and comprehensible direction on how the scores are adjusted in case of a biased rating. To delve into the details and answer research question 3, I mined a number of frequent patterns observed in consumers' rating behavior using the Apriori algorithm which helped to understand the nature of their score adjustment due to bias. Finally, by processing all the information deduced from the results of the statistical analysis and the frequent pattern mining algorithm, I aim to answer research question 4 and offer specific design recommendations. The analyses of the rating behavior are described in Chapter 5.

4.3 Selection of Rating Scales

Studies discussed in Chapter 3 have shown that colors can potentially manipulate how consumers interpret the numerical intervals of rating scales which eventually lead to bias in consumers' rating behavior. In order to investigate such bias in users' behavioral patterns, I selected five different rating scales to implement in the user study which would facilitate the users to provide their assessment of different products. The scales included a baseline scale to collect the actual scores and four experimental scales to portray a bias-inducing environment for capturing the biased rating scores of an individual. The scales were designed using the characteristics described in [79] to represent the scale types defined in [16] in Section 3.2. In this section, I describe the rationale for the selection of the rating scales. The scales used in this research experiment are:

1. Yellow Star Scale
2. Red-Yellow-Green Star Scale
3. Red-Yellow-Blue Star Scale
4. Red-Yellow-Green Emoji Scale
5. Red-Yellow-Blue Emoji Scale

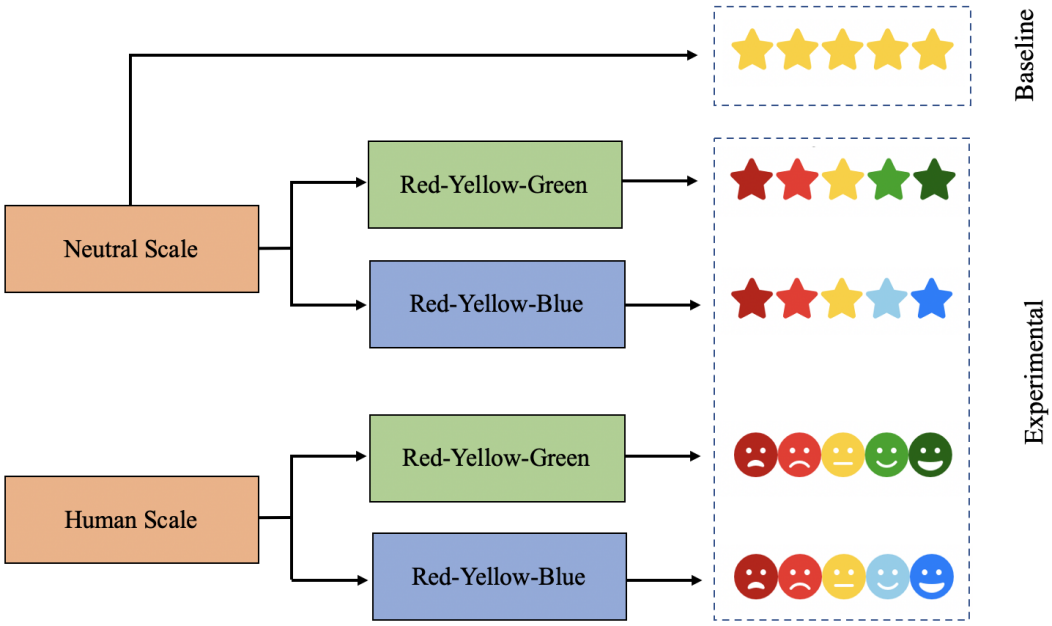


Figure 4.2: Construction of rating scales adopted in the study

4.3.1 Selection of Rating Scale Type

As the first step of scale selection, I decided on the type of rating scales to be employed in the user study. The rating scales chosen for this study are the exemplars of the clusters described by Cena et al. [16]. Among the three clusters, I chose to disregard the "technical scale", because a "technical scale" or slider is the representation of a continuous rating scale which allows the raters to provide a much more granular level of rating scores, while a "human scale" and a "neutral scale" are the representation of typical discrete rating scales [83]. In order to steer clear of the discussion on the difference between continuous and discrete scales, sliders were not included in the set of scales adopted in the study. Another reason for choosing the first two clusters over the third one is consumers' degrees of familiarity with them. While "neutral scales" or stars are the most popular and common, "human scales" or emojis are less popular than stars and "technical scale" or sliders are the least used scales for collecting product evaluation from consumers.

4.3.2 Selection of Rating Scale Characteristics

As the second step, I defined the characteristics of the representatives selected in the previous step. The main concern of this research is to investigate the impact of color metaphors used in the scales on the raters. The biases induced due to the variations in other characteristics of a rating scale namely granularity, labels, neutral point, etc. are beyond the scope of this work. Therefore, scales were designed for the study with no variations

Table 4.1: Brief overview of the features of the chosen rating scales

Features	Scale 1	Scale 2	Scale 3	Scale 4	Scale 5
Neutral Scale	✓	✓	✓		
Human Scale				✓	✓
RYG		✓		✓	
RYB			✓		✓
5-points Granularity	✓	✓	✓	✓	✓
Label					
Neutral Point	✓	✓	✓	✓	✓
RYG = red-yellow-green trio, RYB = red-yellow-blue trio					

in terms of these characteristics. In an attempt to observe the color effect, both the representatives of "human scale" and "neutral scale" are designed using two common color metaphors observed in the literature review. Other than the color and the presentation form, I maintained a conventional and well-established format for all other characteristics so that no other design features could distort the bias resulted from the effect of color. Every scale implemented in this study represented discrete values, had a 5-points granularity and a neutral point; and did not have any negative points, half-scale precision and labels on it. The color metaphors embedded in the scales are mentioned below:

1. **Red-Yellow-Green Color Trio:** The first metaphor embedded in the rating scale is the traffic light metaphor which represents a high contrast color scheme [19]. The color-coded rating scales using the traffic light model has been widely accepted by users [79]. It is also adopted by different platforms such as a food rating system [52], a movie review community called Metacritic, a consumer review website called Trustpilot, etc. According to [75], this color scheme is capable of inducing bias in consumers' rating behavior, since the endpoints are shaded with different hues of two different colors. Researchers also claimed in [13] that the presence of green at the upper endpoint, when paired with red at the other endpoint, can lead to rating score adjustments . All these findings point to the capability of this scale to induce bias and hence this is my first choice of the color trio to apply on the experimental scales.

   Applying this color combination on the "neutral" and the "human" scales produced scales 2 and 4, with the darkest shades of red and green indicating the most extreme negative and positive endpoints of the scales respectively. It is also to be noted that, the two different shades of red represented the low scores (1 and 2), yellow represented the medium score (3) and the two different shades of green represented the high scores (4 and 5).

2. **Red-Yellow-Blue Color Trio:** Another contrasting color scheme used in the scales is the red-yellow-blue trio [36]. The combination of red at the lower endpoint and blue at the upper endpoint of a rating scale has been widely accepted by the respondents from previous research works [57, 75] because the

contrasting color combination makes the endpoints appear more distinctive. In addition, the yellow color is added to the experimental scales in order to make the neutral point distinguishable. Both the color combinations might also induce cognitive bias and consequently cause rating score adjustments among respondents with different personality traits and cultures due to their preferential choices of color.

Applying the color trio on the "neutral" and the "human" scales produced scales 3 and 5, with the darkest shades of red and blue indicating the most extreme negative and positive endpoints of the scales respectively. The two different shades of red represented the low scores (1 and 2), yellow represented the medium score (3) and the two different shades of blue represented the high scores (4 and 5).

4.3.3 Baseline Scale

To represent consumers' ratings in the absence of the color effect, the yellow star rating scale with 5 points of granularity was employed in the user study. The ratings provided by the participants using this scale would be considered as their true ratings or the baseline treatment since it is known as the most popular rating scale and designed with a monochromatic scheme. According to Alghamdi et al., it is the standard for collecting consumers' product evaluation and used by many online review systems i.e. Amazon, IMDb, Booking.com, etc. [7]. In [15], the authors carried out surveys on 45 online rating systems in a variety of domains including movies, travels, etc. and found that the star-based scale with 5 points of granularity was the most exploited rating scale. With the exception of not using any color metaphor, the scale is consistent in other characteristics with the experimental scales. If participants' ratings that were provided using the color-coded scales for the same products statistically significantly differ in average from the baseline score, it would corroborate the claim that emotional connotations carried by the color-coded rating scales can distort the true ratings of consumers depending on their personality and culture. Table 4.1 provides an overview of the features of the scales employed in the user study.

4.4 Design and Implementation of User Study

The study was designed to elicit participants' biased responses to different color-coded rating scales and map their rating behavior to their individuality. The main interest of this research is in the relationship between consumers' bias and their behavioral determinants i.e. personality trait and culture as they might provide directions to designing customized rating scales for consumers with different personalities and cultures. Customized rating scales according to consumers' individuality can contribute significantly to avoid bias in their rating behavior that could be instigated by a one-size-fits-all scale. Therefore, it is important to investigate the relationship between consumers' individuality and their rating bias.

The user study was designed in order to collect responses regarding users' demographics and personality types and to collect their ratings. It was implemented on a web-based platform so that the study is available

and accessible to a large number of diverse participants regardless of their place and time. The user study was developed using HTML5 (Hypertext Mark-up Language), JavaScript and CSS (Cascading Style Sheets). It was deployed on Heroku, which is a free cloud application platform to facilitate developers with the development, deployment and monitoring of applications in the cloud. The online procedure provided the necessary instructions on the demographic questionnaire, the personality assessment questionnaire and the rating activities to make it easier for the participants to complete them. Prior to participating in the study, the participants were first debriefed on the objective and the prerequisite of the experiment and the privacy concerns related to it through an online consent form written by following the guidelines given by the Research Ethics Board of the University of Saskatchewan. As a prerequisite, a participant should have familiarity with the online rating systems. After a participant has given his consent by clicking the "I Agree" button, the step-by-step instructions appear to help him fulfill the required tasks. A participant could leave the study anytime they want. The study contains three main tasks:

1. Completing the demographic questionnaire including age, gender and home country.
2. Completing the Big Five 44-items personality assessment questionnaire.
3. Selecting and rating products using the baseline and the experimental scales.

In the following sections, I describe the aforementioned steps and how I integrated them into the study.

4.4.1 Demographic Survey

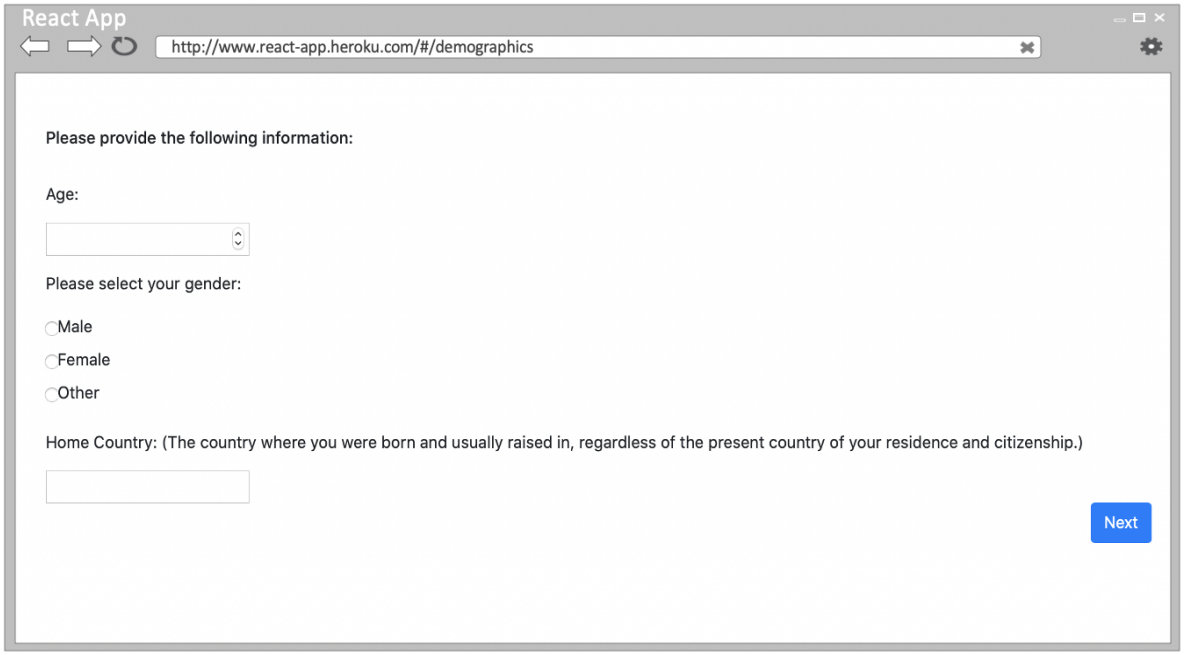


Figure 4.3: Demographic survey of the study

At the initial stage of the study, the participants had to fill in a set of questions about their demographics by indicating their age, gender and home country. While age and gender served the purpose of getting an overall idea about the demographics of the participants, their home country would help decide whether a particular participant belongs to a collectivist or an individualist culture. According to [25], the individualism score of a country indicates the culture practiced in that country. A low score refers to a collectivist society whereas a high score refers to an individualistic society.

4.4.2 Big Five Survey

In order to assess participants' personality traits, they were asked to complete the Big Five Inventory (BFI, 44 items) questionnaire, which is a self-report inventory consisting of short phrases with relatively attainable vocabulary. The assessment requires the participants to answer 44 questions about themselves on a scale of 1-5 (1-Disagree strongly, 2-Disagree a little, 3-Neither agree nor disagree, 4-Agree a little, 5-Agree strongly). The inventory possesses 7-9 items for each trait and it takes 5 minutes to complete the assessment. Based on the answers given by the participants, the score for each trait is calculated following the specified scoring method. The highest score represents the most dominant trait in an individual. The questionnaire and the scoring instructions according to [40] are mentioned below:

1. Is talkative.
2. Tends to find fault with others.
3. Does a thorough job.
4. Is depressed, blue.
5. Is original, comes up with new ideas.
6. Is reserved.
7. Is helpful and unselfish with others.
8. Can be somewhat careless.
9. Is relaxed, handles stress well.
10. Is curious about many different things.
11. Is full of energy.
12. Starts quarrels with others.
13. Is a reliable worker.
14. Can be tense.
15. Is ingenious, a deep thinker.
16. Generates a lot of enthusiasm.
17. Has a forgiving nature.
18. Tends to be disorganized.
19. Worries a lot.
20. Has an active imagination.
21. Tends to be quiet.
22. Is generally trusting.
23. Tends to be lazy.
24. Is emotionally stable, not easily upset.
25. Is inventive.
26. Has an assertive personality.
27. Can be cold and aloof.
28. Perseveres until the task is finished.
29. Can be moody.
30. Values artistic, aesthetic experiences.
31. Is sometimes shy, inhibited.
32. Is considerate and kind to almost everyone.

33. Does things efficiently.
34. Remains calm in tense situations.
35. Prefers work that is routine.
36. Is outgoing, sociable.
37. Is sometimes rude to others.
38. Makes plans and follows through with them.
39. Gets nervous easily.
40. Likes to reflect, play with ideas.
41. Has few artistic interests.
42. Likes to cooperate with others.
43. Is easily distracted.
44. Is sophisticated in art, music, or literature.

In the scoring method, R indicates the reverse-scored items. The originally chosen scale should be subtracted from 6 for every reverse-scored item for each individual. The score is then calculated by adding up the following items for each Big Five trait:

1. Extraversion: 1, 6R, 11, 16, 21R, 26, 31R, 36
2. Agreeableness: 2R, 7, 12R, 17, 22, 27R, 32, 37R, 42
3. Conscientiousness: 3, 8R, 13, 18R, 23R, 28, 33, 38, 43R
4. Neuroticism: 4, 9R, 14, 19, 24R, 29, 34R, 39
5. Openness: 5, 10, 15, 20, 25, 30, 35R, 40, 41R, 44

Once the score for every trait is summed up, a calculation is performed using the following equation to determine an individual's dominant personality trait [42]:

- P : Summation of the points of every item for each trait.
- P_{Min} : For every item, the minimal point is 1. For a trait with n number of items, $P_{Min}= 1*n$
- P_{Max} : For every item, the maximum point is 5. For a trait with n number of items, $P_{Max}= 5*n$
- P_M : For every item, the middle of the scale is 3. For a trait with n number of items, $P_M= 3*n$

$$T = \left(50 + \frac{100}{P_{Max} - P_{Min}}(P - P_M)\right)[\%] \tag{4.1}$$

The resultant score T obtained from the equation refers to the percentage of a given personality trait. If the score is less than 50%, it is considered as the score of the opposite trait and is measured as $(100 - T)\%$. For example, if a subject has scored 30% in Agreeableness, then it is considered that he has scored 70% in Antagonism. According to [82], participants can be categorized based on their extremely dominant, moderately dominant and minimally dominant personality traits. For a participant, the score for every trait is calculated first. Next, to observe their personality-based rating behavior, I classified the subjects into their extremely dominant personality groups.

4.4.3 User Rating Collection

Upon giving their consent to participate in the study, the subjects were first presented with the demographic questionnaire. After the completion of it, they were asked to choose the items they had used or consumed before from a list of 21 products. The products were chosen from 7 different domains based on their popularity and the consumers' familiarity with them. For example, from the domain of social networking sites, I selected the trendiest 3 sites: Facebook, Twitter and LinkedIn [22] because they are among the top social sites used around the world. The process of collecting user ratings on the products was the combination of two phases: a baseline phase and an experimental phase.

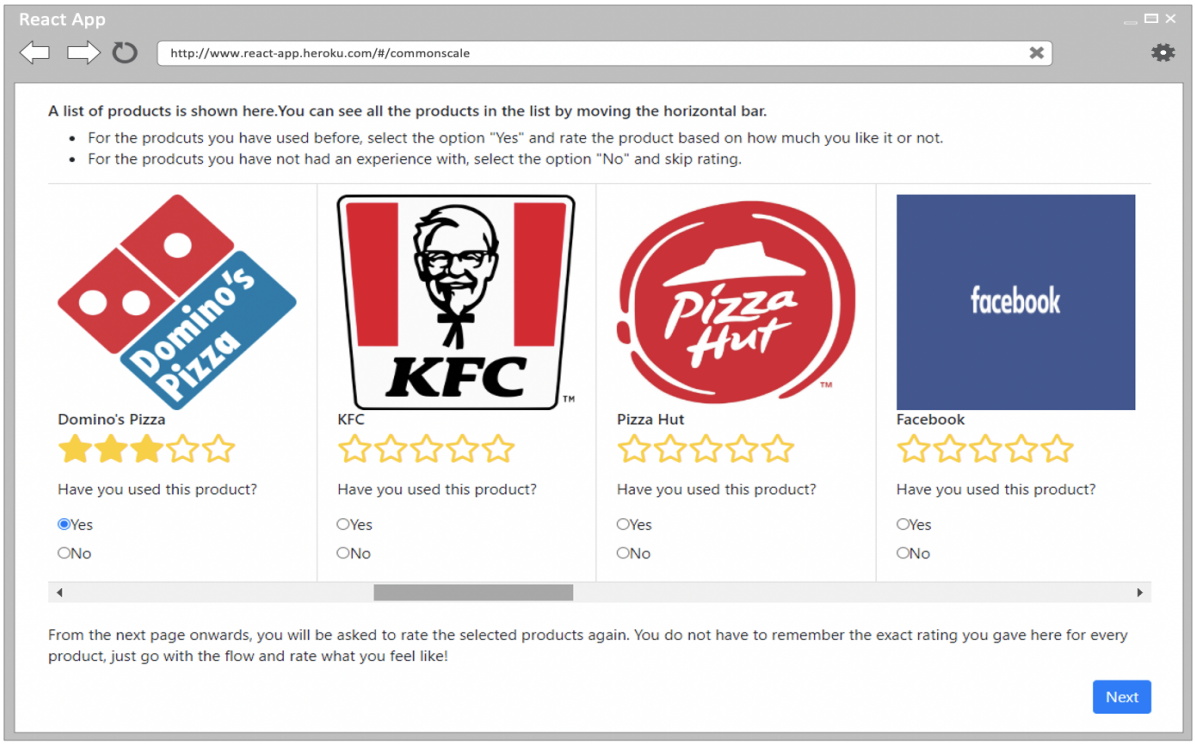


Figure 4.4: Baseline Phase: Rating collection using the baseline scale

In the baseline phase, participants selected the products they had already experienced and provided their post-consumption feedback once on every selected product using the baseline scale. They also selected an answer to the question "Have you used this product?" between the options "yes" or "no" for each product to ensure that the raters did not rate any product they had not used before. The ratings provided in this phase mirrored their genuine evaluation scores since the baseline scale represented the absence of any color effect and was designed using a yellow monochromatic scheme.

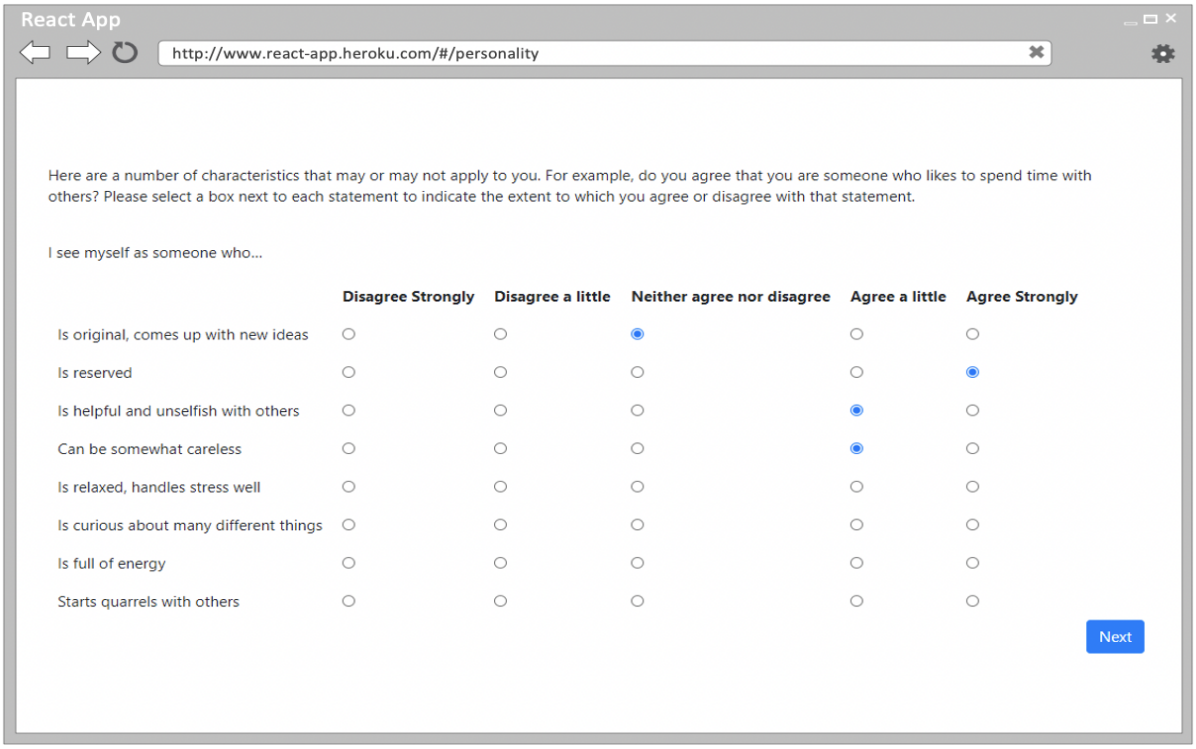


Figure 4.5: Experimental Phase: A segment of the Big Five survey

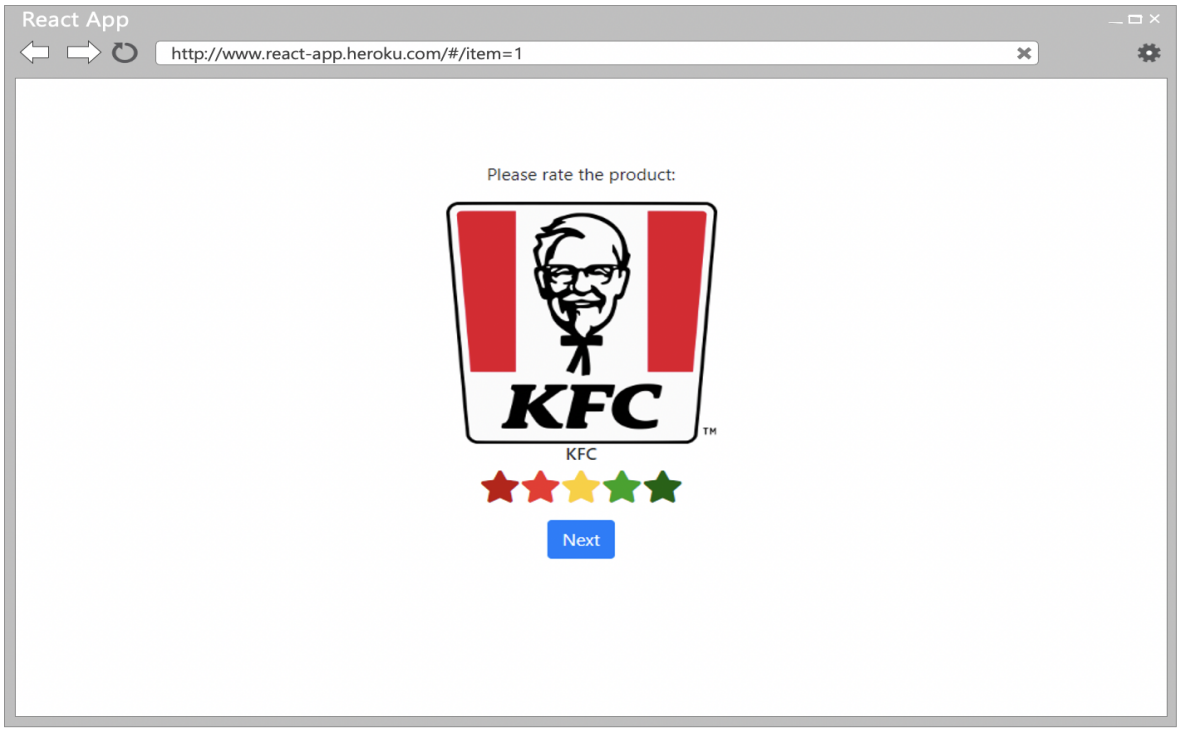


Figure 4.6: Experimental Phase: Rating collection in the red-yellow-green star-based scale

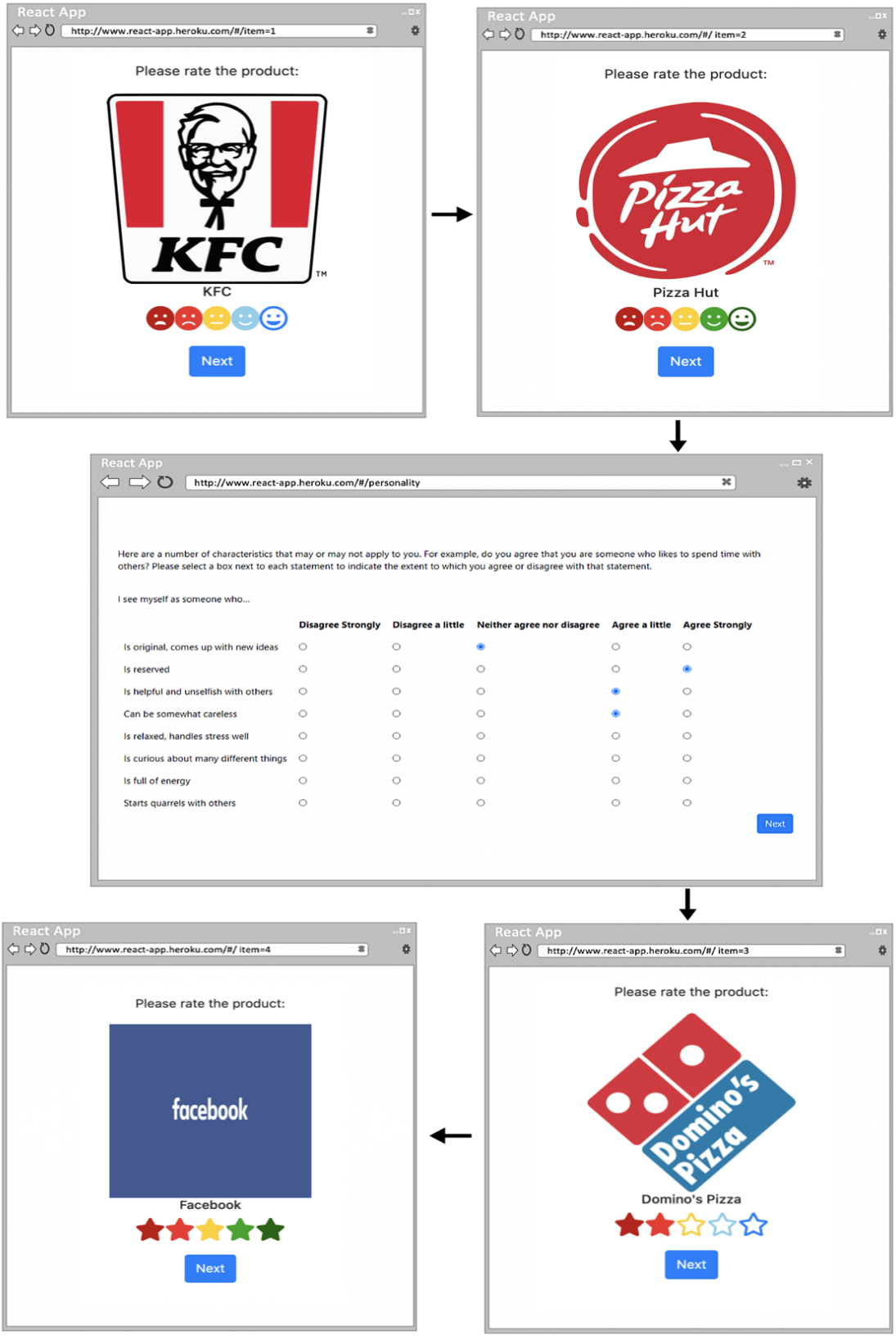


Figure 4.7: Flow of the Experimental Phase: A segment of the rating collection process with randomized sequence of rating activities and Big Five survey

In the experimental phase, participants completed the Big Five survey (as shown in Figure 4.5) and provided their ratings again on the products they selected in the baseline phase using the four experimental scales (Figure 4.6). In an experiment that is designed with several conditions, participants can be assigned to the conditions in two ways: between-subjects and within-subjects designs. A between-subjects experimental design refers to the design where a participant is allocated to any one of the conditions and a within-subjects study is where a participant is engaged in all conditions [65]. Since my research follows a within-subject design, order effects could occur in this context. According to [71], order effects refer to the impact on the participants' responses resulting from the influence of the order in which the conditions are presented in the survey. I counterbalanced the order effects by randomizing not only the order of appearance of the experimental scales but also the order of items to rate. Furthermore, the personality assessment survey was split into several sections and incorporated with the experimental phase of user rating collection so that the rating activities would not become monotonous and the participants would not grow accustomed to the scores they provided for the products.

As shown in Figure 4.8, after the completion of the study, participants were also invited to optionally provide their email address to be entered into a raffle for one of the three gift vouchers valued at 50 CAD. If they did not wish to enter into the raffle, they might proceed and finish the survey without providing their email address.

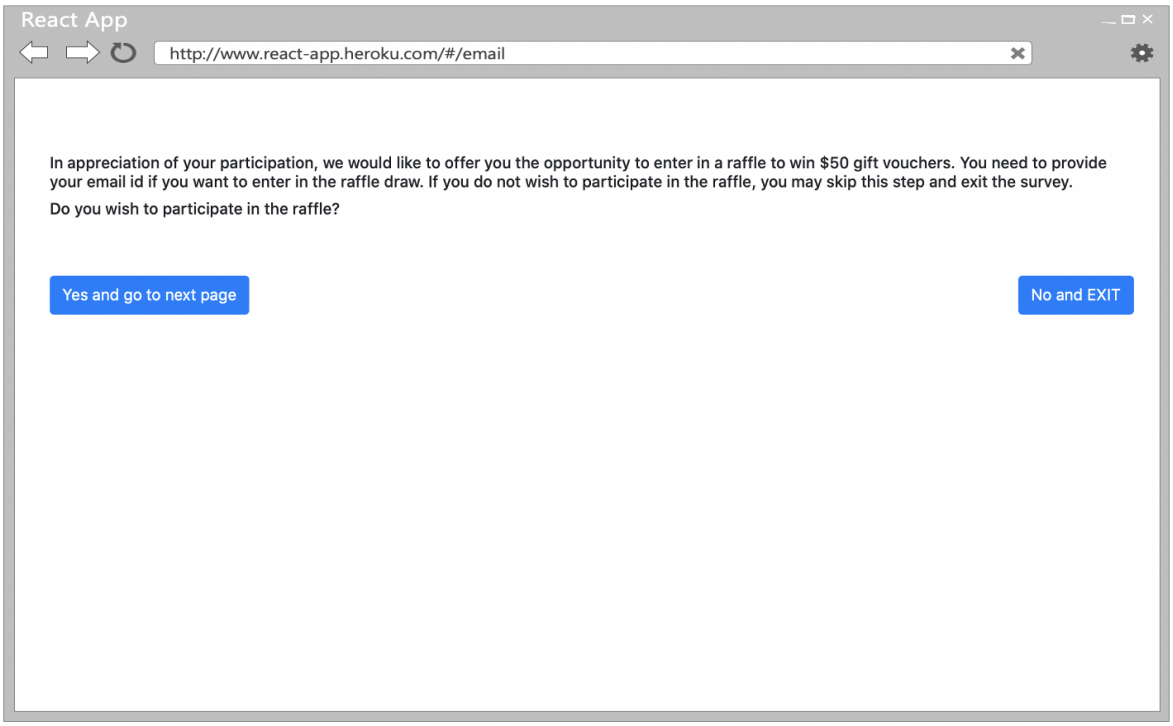


Figure 4.8: Invitation to optionally provide an email address to be entered into a raffle

5 Rating Behavior Analysis

The goal of this analysis is to examine whether consumers utilize different color-coded rating scales differently for the same product on the basis of their individuality. According to [70] and [76], personality and culture are two stable aspects to shape the individual behavioral pattern, hence I investigate and analyze the influence of these aspects on consumers' rating behavior across different color-coded rating scales. The analysis conducted in this research involves two approaches. Firstly, I conducted a statistical analysis by utilizing the Wilcoxon signed rank test to understand whether a statistically significant difference appears in the rating scores assigned on different scales by the participants with different personalities and cultures. The second approach entails association rule mining with the Apriori algorithm to capture the direction of their rating score adjustments in cases where the consumers exhibited biased rating behavior. The following sections present the results obtained from the data analysis, discusses the key findings and answers the research questions raised in Section 4.1.

5.1 Participants

Recruitment for the study was accomplished through announcements made on PAWS which is a web environment provided by the University of Saskatchewan and through advertisements posted on social platforms including Facebook and LinkedIn. The data collection process lasted for a period of five months (from April 2020 to August 2020). Initially, a total of 192 subjects' complete responses were collected but eventually, not all of them were considered for the final analysis. To evaluate the subjects' eligibility to be included in the final dataset, I assessed the reliability of their BFI responses. Among 44 items, the Big Five Inventory possesses 16 pairs of items with opposite implications of a personality and the reliability assessment is conducted by inspecting the consistency of these pairs [35]. I excluded the subjects who exhibited more than 4 inconsistencies out of 16 pairs of items and consequently the final database was left with 176 participants with complete and reliable responses.

Table 5.1 shows the demographics of the participants: 64 males (36.4%) and 110 females (62.5%). All participants were at least 17 years old. The majority (56.8%) of the participants aged between 21 to 30 years, 22.7% aged below 21 years and participants above 30 years were the minorities (20.5%). Of all the participants, two main nationalities were Bangladeshi (48 participants) and Canadian (73 participants), the rest of them were from different countries including China, the United States, Nigeria, Brazil, etc.

Table 5.1: Participants' demographics

Criterion	Group	Bangladesh	Canada	Other Countries	Total	Percentage
Gender	Male	20	15	29	64	36.4%
	Female	28	57	25	110	62.5%
	Other	0	1	1	2	1.1%
Age	<21 years	7	21	12	40	22.7%
	21-30 years	39	37	24	100	56.8%
	>30 years	2	15	19	36	20.5%

5.2 Personality-wise Rating Behavior Analysis

To observe personality-wise rating behavior across different color-coded rating scales, respondents were first classified into their traits. According to the personality acquisition test, 40 subjects were classified into the group of Openness to experience, 30 subjects were classified into the group of Conscientiousness, 28 subjects were in the group of Extraversion, 42 subjects in Agreeableness trait, 23 subjects in Neuroticism, 11 subjects were in the group of Introversion (the opposite trait of Extraversion) and 5 subjects were in the group of Emotional stability (the opposite trait of Neuroticism). However, no subject was categorized into the groups of Closeness to experience (the opposite trait of Openness to experience), Lack of direction (the opposite trait of Conscientiousness) and Antagonism (the opposite trait of Agreeableness).

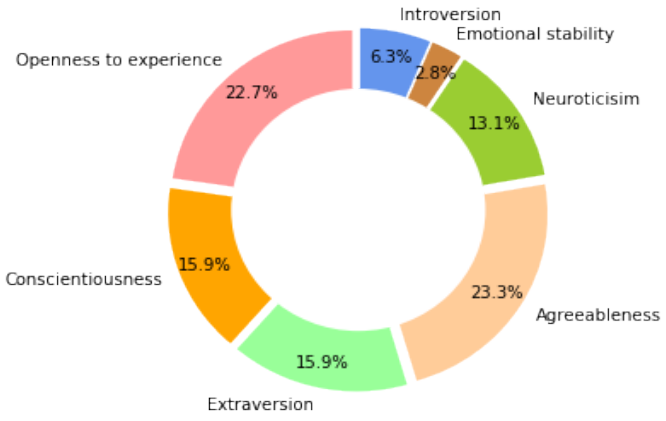


Figure 5.1: Percentage of participants grouped by personality traits

To get an overall understanding of the measure of central tendency and the measure of dispersion, the descriptive statistics of the ratings given by each personality group in the baseline and all four experimental scales are shown in the following figures.

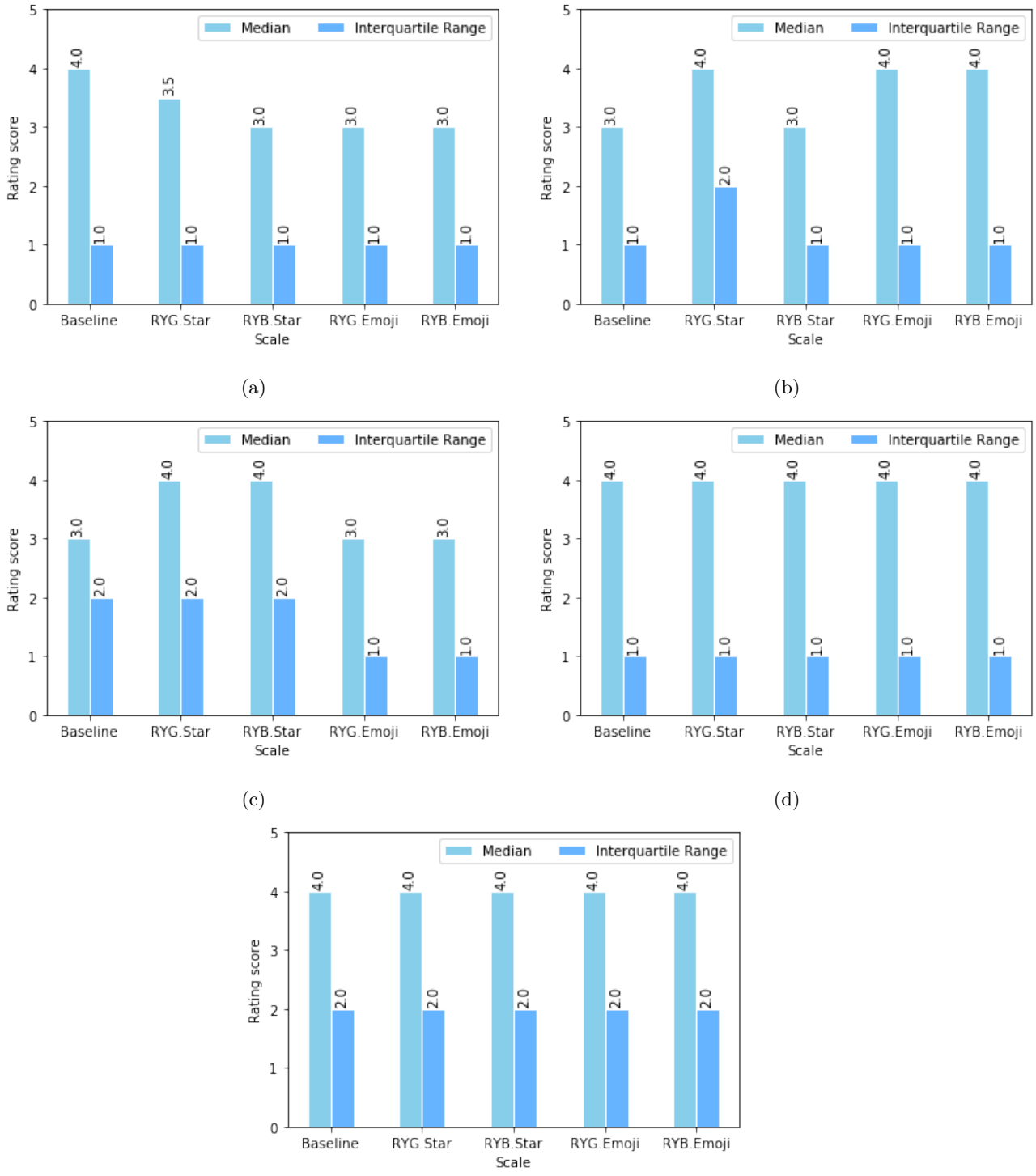
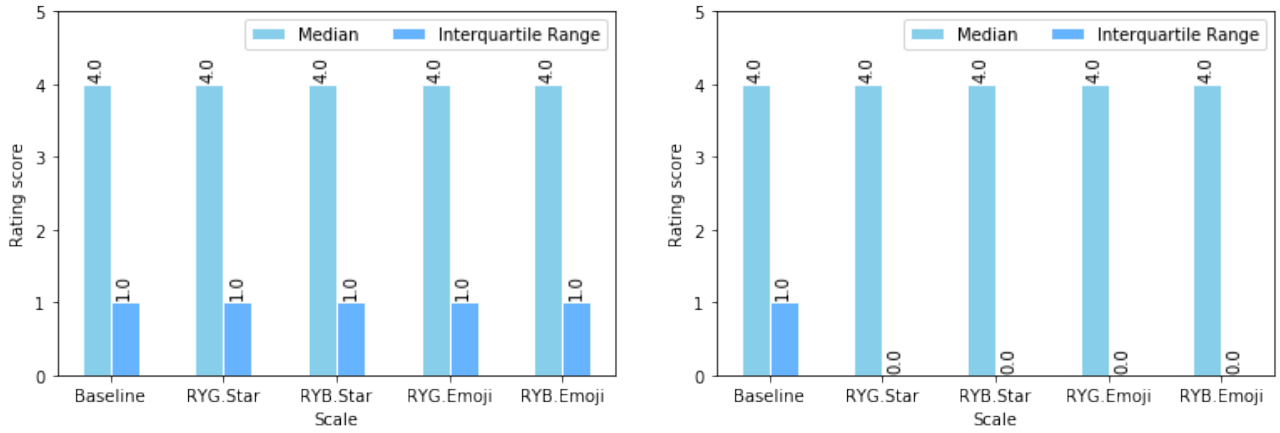


Figure 5.2: Descriptive statistics of user ratings grouped by the traits (a)Openness to experience (b)Conscientiousness (c)Extraversion (d)Agreeableness and (e)Neuroticism.



(a) (b)

Figure 5.3: Descriptive statistics of user ratings grouped by the traits (a)Introversion and (b)Emotional stability.

According to the descriptive statistics, the median (Mdn) is the measure of central tendency of a distribution that represents the middle of that distribution. And the interquartile range (IQR) represents the measure of dispersion that shows whether the responses of a distribution are consensus or not. Five rating scales are represented in the plots of descriptive statistics as: Baseline (yellow star scale), RYG.Star (red-yellow-green star scale), RYB.Star (red-yellow-blue star scale), RYG.Emoji (red-yellow-green emoji scale) and RYB.Emoji (red-yellow-blue emoji scale). According to Figure 5.2(a), for the participants with high openness to experience, although the average score in the RYG.Star (Mdn=3.5, IQR=1), RYB.Star (Mdn=3, IQR=1), RYG.Emoji (Mdn=3, IQR=1) and RYB.Emoji (Mdn=3, IQR=1) scales are dissimilar to the baseline (Mdn=4, IQR=1), the scores in all five scales have a similar variability around the median. For respondents with high conscientiousness, the average score in every experimental scale is dissimilar (RYG.Star (Mdn=4, IQR=2), RYG.Emoji (Mdn=4, IQR=1) and RYB.Emoji (Mdn=4, IQR=1)) to the baseline (Mdn=3, IQR=1), except for RYB.Star scale (Mdn=3, IQR=1) but the scores were more widely distributed in RYG.Star scale than others. As depicted by Figure 5.2(c), the average of the rating scores given by extroverts in the baseline (Mdn=3, IQR=2) was similar to that in RYG.Emoji (Mdn=3, IQR=1) and RYB.Emoji (Mdn=3, IQR=1) scales, but lower than that in RYG.Star (Mdn=4, IQR=2) and RYB.Star (Mdn=4, IQR=2) scales. The scores were also more widely distributed in the star-based scales than others. As shown in Figure 5.2(d), the median of the ratings of the agreeable subjects was equal and the scores had a similar variance across every scale (for each scale, Mdn=4, IQR=1). Similar to agreeable participants, neurotics had an indifferent average score in every scale and their scores were similarly distributed as well(for each scale, Mdn=4, IQR=2).

On the other hand, Figure 5.3 depicts the overall summary of the ratings given by the participants who are introverts or emotionally stable. For introverts, no difference was observed in either the average or the

variance of the distribution of scores across the scales (for each scale, Mdn=4, IQR=1). Lastly, in the case of emotionally stable participants, the median was the same across the scales but the scores were more widely distributed in the baseline (Mdn=4, IQR=1) than the experimental scales (for each of the four experimental scales, Mdn=4, IQR=0).

Table 5.2: Wilcoxon signed rank test: Results grouped by the personality traits

Trait	Pairwise comparison	Z	p-value (2-tailed)
Openness to experience	Baseline and RYG.Star	-0.026	0.979
	Baseline and RYB.Star	-0.155	0.877
	Baseline and RYG.Emoji	-0.803	0.422
	Baseline and RYB.Emoji	-0.265	0.791
Conscientiousness	Baseline and RYG.Star	-0.700	0.484
	Baseline and RYB.Star	-1.077	0.282
	Baseline and RYG.Emoji	-0.022	0.982
	Baseline and RYB.Emoji	-0.291	0.771
Extraversion	Baseline and RYG.Star	**-4.092**	**0.000**
	Baseline and RYB.Star	**-4.242**	**0.000**
	Baseline and RYG.Emoji	-1.485	0.137
	Baseline and RYB.Emoji	-0.231	0.817
Agreeableness	Baseline and RYG.Star	-1.051	0.293
	Baseline and RYB.Star	-1.844	0.065
	Baseline and RYG.Emoji	-1.437	0.151
	Baseline and RYB.Emoji	-1.604	0.109
Neuroticism	Baseline and RYG.Star	-0.404	0.686
	Baseline and RYB.Star	-0.317	0.752
	Baseline and RYG.Emoji	-0.605	0.545
	Baseline and RYB.Emoji	-0.179	0.858
Introversion	Baseline and RYG.Star	-0.592	0.554
	Baseline and RYB.Star	-0.517	0.605
	Baseline and RYG.Emoji	-0.354	0.723
	Baseline and RYB.Emoji	-0.577	0.564
Emotional stability	Baseline and RYG.Star	-1.713	0.087
	Baseline and RYB.Star	-1.030	0.303
	Baseline and RYG.Emoji	-0.617	0.537
	Baseline and RYB.Emoji	-0.421	0.674
The bold values are indicating to a statistically significant difference ($p \leq 0.05$).			

Table 5.3: Rank statistics for the users with high extraversion in Baseline and RYG.Star scales

		N	Mean rank	Sum of ranks
Ratings in RYG.Star-Ratings in Baseline	Negative ranks	38[a]	62.22	2364.50
	Positive ranks	88[b]	64.05	5636.50
	Ties	110[c]		
	Total	236		

a. Ratings in RYG.Star<Ratings in Baseline

b. Ratings in RYG.Star>Ratings in Baseline

c. Ratings in RYG.Star=ratings in Baseline

Table 5.4: Rank statistics for the users with high extraversion in Baseline and RYB.Star scales

		N	Mean rank	Sum of ranks
Ratings in RYB.Star-Ratings in Baseline	Negative ranks	39[a]	64.77	2526.00
	Positive ranks	92[b]	66.52	6120.00
	Ties	115[c]		
	Total	236		

a. Ratings in RYB.Star<Ratings in Baseline

b. Ratings in RYB.Star>Ratings in Baseline

c. Ratings in RYB.Star=ratings in Baseline

The descriptive statistics, so far, summarized the characteristics and helped to gain an overall understanding of the ratings assigned by each personality group. However, to evaluate the potential significant difference among the user ratings given in all five scales by each personality group, I validated the data for the Wilcoxon assumptions and performed the Wilcoxon signed rank test. It is a non-parametric statistical method for pairwise comparison, employed using the personality trait as the within-subject factor. Prior to that, I performed the Shapiro-Wilk test to evaluate the normality of the distribution of the differences between the two related groups (i.e. the distribution of differences between the rating scores of the baseline scale and red-yellow-green star scale for the users with high agreeableness). The Shapiro-Wilk test showed a significant deviation from normality ($p=0.000$, all p's<0.05) for each paired group which validated the data for the non-parametric assumptions.

The Wilcoxon signed rank test compared between the ratings obtained from the pairwise combination of the baseline and each one of the four experimental scales. Each trait, therefore, is comprised of four groups (i.e. Baseline and RYG.Star, Baseline and RYB.Star, Baseline and RYG.Emoji, Baseline and RYB.Emoji scale). The results of the inferential statistics for the participants of the personality traits are presented in Table 5.2. Along with the statistical significance, I also measured the effect size to quantify the difference

between the paired samples. The analysis revealed that the ratings provided by the extroverts using RYG.Star scale (Z= -4.092, p= 0.000) are statistically significantly different from their ratings in the baseline scale with a small effect size, r= 0.19. It evidently confirmed that extroverts' original post-consumption ratings were distorted under the influence of RYG.Star scale. Furthermore, the result of the analysis proved the existence of a statistically significant difference between the ratings given using RYB.Star scale and baseline scale (Z= -4.242, p= 0.000) with a small effect size, r= 0.20. Therefore, a rating bias was also apparent in the way extroverts utilized the RYB.Star scale. However, no statistically significant difference was found in the rating behavior of any other group of participants (i.e. openness to experience, conscientiousness, agreeableness, neuroticism, introversion and emotional stability).

Table 5.3 and Table 5.4 provide detailed information about the ranked scores for extroverts in the RYG.Star scale and the RYB.Star scale respectively. According to Table 5.3, the negative ranks indicate the ranks where the rating score in the baseline was higher than the score in RYG.Star scale, the positive ranks indicate the ratings which are higher in RYG.Star scale than the baseline scale and tied ranks indicate cases where the ratings were indifferent. The table shows that for 88 of the 236 cases, the ranks were positive, indicating extroverts' tendency to assign higher score in RYG.Star scale than the baseline. Whereas in 38 cases only, the ratings in the RYG.Star scale were lower. In total, in 126 out of 236 cases, the ratings were different, confirming the existence of bias in extroverts' rating behavior regardless of the direction the biased ratings were adjusted to. Whereas in 110 cases, the scores had tied ranks.

Similarly, from Table 5.4 it can be inferred that in 92 cases of positive ranks out of 236 cases, indicating to extroverts' tendency to rate higher in RYB.Star scale than the baseline scale. In 39 out of 236 cases the ranks were negative, meaning that the ratings were lower in RYB.Star scale than the baseline. In 115 cases, extroverts' rating scores between the scales did not differ. In both the star-based scales, the mean and the sum of positive ranks are significantly higher than negative ranks.

The statistical analysis, so far, showed results that extroverts would adjust their rating scores in some specific color-coded scales. Yet, it did not investigate the score that is being subject to the bias and the direction of score adjustment due to the bias. To address the gap, I employed the Apriori algorithm for extroverts to mine their rating patterns. As an initial requirement for finding interesting association rules, the value of minimum support was set to 0.1, minimum confidence to 0.75 and minimum lift to 1.0. The candidate rules which did not satisfy the threshold marked by the parameters were discarded. The procedure produced 61 frequent itemsets and 142 association rules for extroverts. Each of the rules can be perceived as an implication of how they are utilizing different scales to rate the same products. The outcomes of the analysis on the rating behavior of the extraversion group are presented in Table 5.5. For each of the five scales, the rating scores are discretized in three distinct forms of representations of a low, a neutral and a high rating score as {1,2},{3} and {4,5} respectively, preceded by a rating scale X, where X={Baseline, RYG.Star, RYB.Star, RYG.Emoji, RYB.Emoji}. For instance, Baseline={1,2} represents a low rating, Baseline={3} represents a neutral score and Baseline={4,5} represents a high score given by the rater in the baseline scale.

Table 5.5: 20 notable association rules for extroverts

No.	LHS	RHS	Support	Confidence	Lift
1	{Baseline = {4,5}, RYG.Emoji = {4,5}, RYB.Emoji = {4,5}}	{RYB.Star = {4,5}}	0.30	0.98	1.69
2	{Baseline = {4,5}, RYB.Emoji = {4,5}, RYB.Star = {4,5}}	{RYG.Emoji = {4,5}}	0.30	0.98	2.07
3	{Baseline = {4,5}, RYB.Star = {4,5}}	{RYG.Star = {4,5}}	0.32	0.97	1.66
4	{Baseline = {4,5}, RYG.Emoji = {4,5}}	{RYB.Star = {4,5}}	0.31	0.97	1.67
5	{Baseline = {4,5}, RYG.Emoji = {4,5}, RYB.Star = {4,5}}	{RYB.Emoji = {4,5}}	0.30	0.97	2.31
6	{Baseline = {4,5}, RYG.Star = {4,5}}	{RYB.Star = {4,5}}	0.32	0.93	1.61
7	{Baseline = {1,2}, RYG.Star = {1,2}, RYB.Emoji = {1,2}}	{RYG.Emoji = {1,2}}	0.10	0.92	4.53
8	{Baseline = {3}, RYG.Star = {4,5}, RYB.Emoji = {3}}	{RYB.Star = {4,5}}	0.10	0.92	1.59
9	{Baseline = {3}, RYG.Star = {4,5}}	{RYB.Star = {4,5}}	0.15	0.90	1.55
10	{RYG.Star = {3}, RYG.Emoji = {3}}	{RYB.Star = {3}}	0.15	0.90	4.00
11	{RYG.Star = {3}, RYG.Emoji = {3}}	{RYB.Emoji = {3}}	0.15	0.90	2.65
12	{Baseline = {1,2}, RYG.Star = {1,2}, RYG.Emoji = {1,2}}	{RYB.Star = {1,2}}	0.10	0.88	4.56
13	{Baseline = {1,2}, RYG.Emoji = {1,2}, RYB.Star = {1,2}}	{RYG.Star = {1,2}}	0.10	0.88	4.56
14	{Baseline = {1,2}, RYG.Star = {1,2}, RYG.Emoji = {1,2}}	{RYB.Emoji = {1,2}}	0.10	0.88	3.68
15	{Baseline = {1,2}, RYG.Emoji = {1,2}}	{RYB.Emoji = {1,2}}	0.12	0.88	3.65
16	{Baseline = {1,2}, RYB.Star = {1,2}}	{RYB.Emoji = {1,2}}	0.12	0.85	3.53
17	{Baseline = {3}, RYB.Star = {4,5}}	{RYG.Star = {4,5}}	0.15	0.84	1.43
18	{RYB.Emoji = {3}, RYB.Star = {3}}	{RYG.Star = {3}}	0.14	0.83	3.78
19	{RYB.Star = {3}}	{RYG.Emoji = {3}}	0.18	0.83	2.57
20	{Baseline = {3}, RYG.Emoji = {3}}	{RYB.Emoji = {3}}	0.14	0.82	2.44

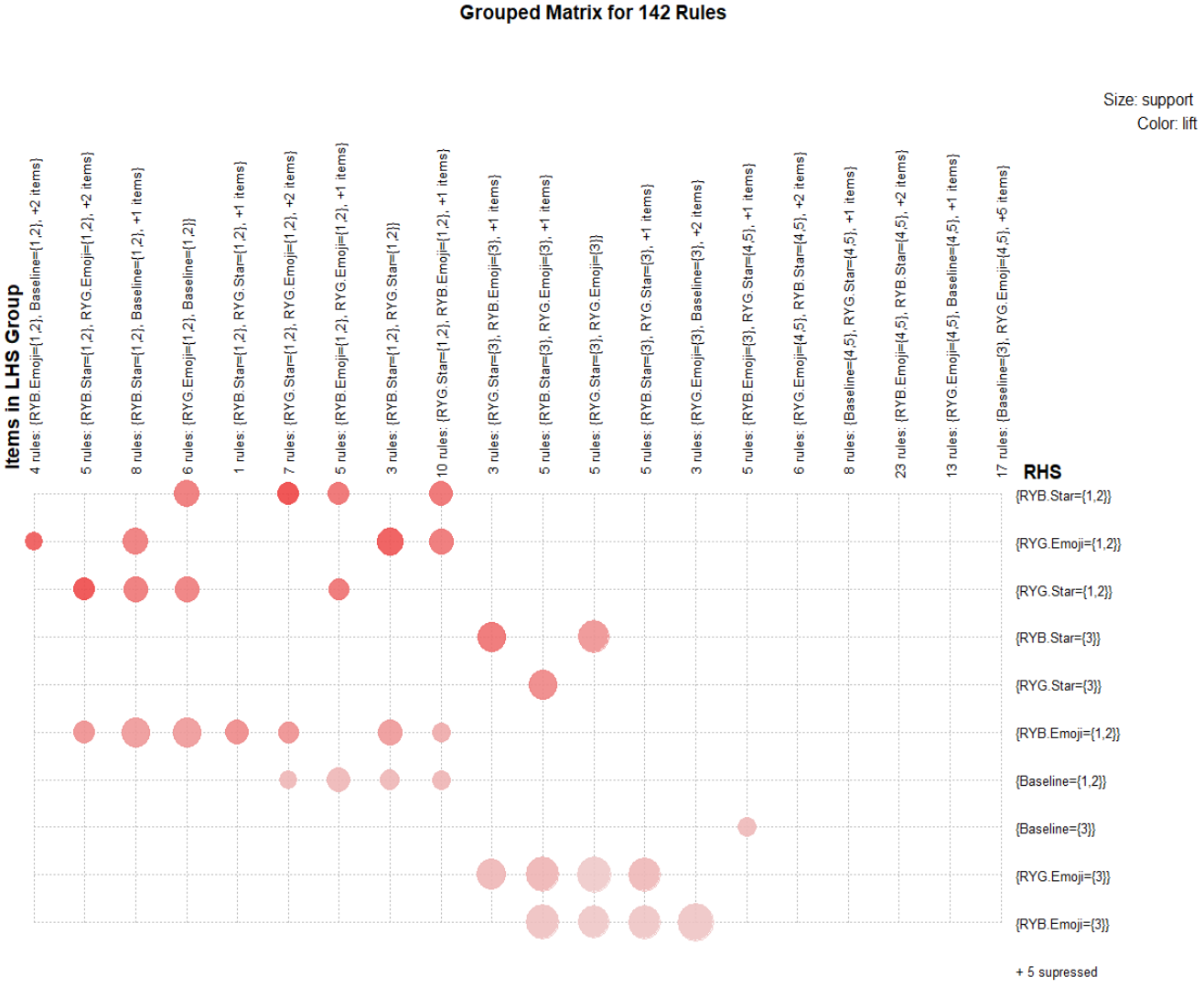


Figure 5.4: Impact of color-coded rating scales on extroverts (rules with confidence ≥ 0.75)

In Table 5.5, 20 notably interesting rules generated from the Apriori algorithm are mentioned which represent the particular direction of score adjustment resulting from the bias of the participants with high extraversion trait. A number of rules captured the evident rating bias in extroverts, with respect to the behavioral differences in their ratings between the baseline scale and the RYG.Star scale, and between the baseline scale and RYB.Star scale. To explain, according to rule 9, if an extrovert has given a neutral score in baseline and a high score in RYG.Star scale, it also means that he will adjust his true rating and assign a high score to the product in RYB.Star scale. Rule 17 depicts a similar pattern to rule 9, that is, when an extrovert consumer has given a high rating in RYB.Star scale and a neutral rating in baseline scale, he will give that product a high rating in RYG.Star scale. Nevertheless, it is evident from rule 20 that such an adjustment is only observable in the color-coded star scales. As stated by rule 20, when a consumer has rated neutrally in

both the baseline scale and RYG.Emoji scale, it also infers that he will not be biased and provide a neutral rating in RYB.Emoji scale as well. Rule 12 to 16 demonstrate that in cases where a product was originally given a low rating, color-coded scales had no impact on users' rating behavior. For example, rule 12 states that, if an extrovert gave a product a low rating in the baseline, RYG.Star and RYG.Emoji scales, it also implies that he will give a low rating in RYB.Star scale. As documented by the first 6 rules, the color-coded scales also failed to distort a high rating score. For instance, rule 3 refers to the fact that when a consumer originally has provided a high rating and also given a high score in RYB.Star scale, it infers that he will give the same rating score using RYG.Star scale.

The grouped matrix-based visualization of the association rules for extroverts is presented in Figure 5.4. The antecedents (LHS) of the 142 rules are plotted along the horizontal axis and the consequents (RHS) are plotted along the vertical axis. The size and the color of the bubble indicate the support and the lift respectively. The bigger the bubble, the greater the support is and the darker the bubble, the higher the lift is. The most interesting group consists of 4 rules which contain "RYB.Emoji={1,2}," and "Baseline={1,2}" and two additional items in the antecedent and the consequent is "RYG.Emoji={1,2}". In summary, the rules indicate that RYG.Star and RYB.Star scales impacted extroverts' rating behavior and to be more specific, the neutral rating scores in the baseline are generally adjusted to the higher endpoints of the RYG.Star and RYB.Star scales due to the bias.

5.3 Cross-cultural Rating Behavior Analysis

In order to discover the cross-cultural rating behavioral pattern across different color-coded rating scales, participants were first segregated into the dimension of collectivist and individualistic culture according to their countries, using Hofstede's individualism score. For example, Bangladesh, with a score of 20 is characterized as a collectivist society whereas Canada, with a score of 80 can be considered an individualistic society. Of 176 participants, 99 subjects were categorized as collectivists and 77 subjects as individualists.

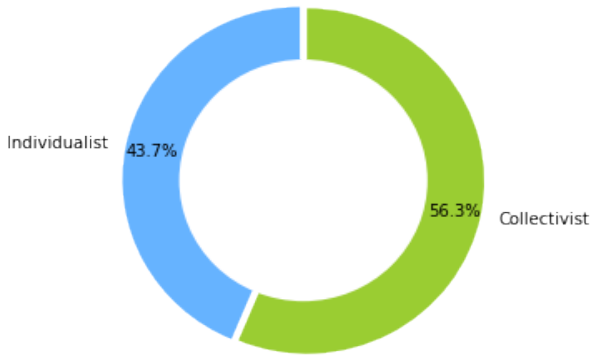


Figure 5.5: Percentage of participants by culture

The descriptive statistics of the ratings given by the collectivists and individualists in different scales shown in Figure 5.6 exhibit their overall rating patterns. Five rating scales are represented as: Baseline (yellow star scale), RYG.Star (red-yellow-green star scale), RYB.Star (red-yellow-blue star scale), RYG.Emoji (red-yellow-green emoji scale) and RYB.Emoji (red-yellow-blue emoji scale). The average rating score of collectivists was indifferent in every scale but the scores in RYG.Star (Mdn=4, IQR=2), RYB.Star (Mdn=4, IQR=2) and RYB.Emoji (Mdn=4, IQR=2) scales were more widely distributed around the median than the baseline (Mdn=4, IQR=1) and RYG.Emoji (Mdn=4, IQR=1) scales. For indivdualists, the median of the scores in the baseline (Mdn=3, IQR=1) was similar to that in RYG.Star (Mdn=3, IQR=1) and RYB.Star (Mdn=3, IQR=1) scales, but lower than RYG.Emoji (Mdn=4, IQR=1) and RYB.Emoji (Mdn=4, IQR=1) scales. However, the scores were similarly distributed around the median across all five scales.

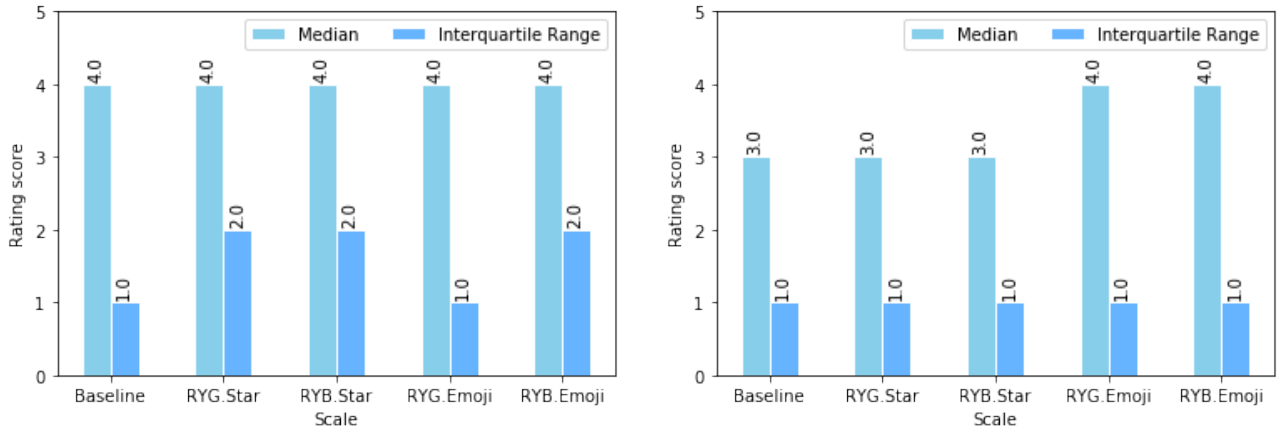


Figure 5.6: Comparison among the descriptive statistics of user ratings grouped by culture (a)Collectivism and (b)Individualism.

The descriptive statistics summarized the characteristics and helped to gain an overview of the ratings assigned by each cultural group. To find out whether the differences between the ratings in the baseline and the four experimental scales are statistically significant or not, I validated the data for the non-parametric assumptions and evaluated the ratings with the Wilcoxon signed rank test, using the collectivist and the individualist cultures as the within-subject factors. The Shapiro-Wilk test conducted to assess the distribution of the differences between the scores for each paired group showed a significant deviation from normality ($p=0.000$, all p's<0.05).

The Wilcoxon signed rank test compared between the rating scores acquired from the paired combination of the baseline and each one of the four experimental scales. Each of the two cultures, is therefore, comprised of four groups (i.e. Baseline and RYG.Star, Baseline and RYB.Star, Baseline and RYG.Emoji, Baseline and RYB.Emoji). The results of the statistical analysis of the ratings assigned by the collectivists and the individualists are shown in Table 5.6.

Table 5.6: Wilcoxon signed rank test: Results grouped by culture

Culture	Pairwise comparison	Z	p-value (2-tailed)
Collectivism	Baseline and RYG.Star	**-3.824**	**0.000**
	Baseline and RYB.Star	**-4.186**	**0.000**
	Baseline and RYG.Emoji	**-2.016**	**0.044**
	Baseline and RYB.Emoji	**-2.369**	**0.018**
Individualism	Baseline and RYG.Star	-0.791	0.429
	Baseline and RYB.Star	-0.988	0.323
	Baseline and RYG.Emoji	-0.062	0.951
	Baseline and RYB.Emoj	-0.639	0.523
The bold values are indicating to a statistically significant difference ($p \leq 0.05$).			

Table 5.7: Rank statistics for collectivists in Baseline and RYG.Star scales

		N	Mean rank	Sum of ranks
Ratings in RYG.Star-Ratings in Baseline	Negative ranks	149^a	174.53	26004.50
	Positive ranks	216^b	188.84	40790.50
	Ties	690^c		
	Total	1055		

a. Ratings in RYG.Star<Ratings in Baseline

b. Ratings in RYG.Star>Ratings in Baseline

c. Ratings in RYG.Star=Ratings in Baseline

Table 5.8: Rank statistics for collectivists in Baseline and RYB.Star scales

		N	Mean rank	Sum of ranks
Ratings in RYB.Star-Ratings in Baseline	Negative ranks	149^a	182.25	27155.50
	Positive ranks	228^b	193.41	44097.50
	Ties	678^c		
	Total	1055		

a. Ratings in RYB.Star<Ratings in Baseline

b. Ratings in RYB.Star>Ratings in Baseline

c. Ratings in RYB.Star=Ratings in Baseline

The pairwise comparison in Table 5.6 revealed a significant impact of the color-coded scales on collectivists. The ratings provided by collectivists using RYG.Star scale are statistically significantly different from their

corresponding ratings in the baseline scale (Z=-3.824, p=0.000) with a small effect size, r= 0.1. The ratings assigned by them in RYB.Star scale are also statistically significantly different from the baseline (Z=-4.186, p=0.000) with a small effect size, r= 0.1. Furthermore, there is a significant difference between the baseline and RYG.Emoji scales (Z=-2.016, p=0.044 and effect size, r= 0.04) and also between the baseline and RYB.Emoji scales (Z=-2.369, p=0.018 and effect size, r= 0.05). There is no significant difference in the rating behavior of the individualists across the scales.

Table 5.9: Rank statistics for collectivists in Baseline and RYG.Emoji scales

		N	Mean rank	Sum of ranks
Ratings in RYG.Emoji-Ratings in Baseline	Negative ranks	160[a]	184.77	29562.50
	Positive ranks	205[b]	181.62	37231.50
	Ties	690[c]		
	Total	1055		

a. Ratings in RYG.Emoji<Ratings in Baseline

b. Ratings in RYG.Emoji>Ratings in Baseline

c. Ratings in RYG.Emoji=Ratings in Baseline

Table 5.10: Rank statistics for collectivists in Baseline and RYB.Emoji scales

		N	Mean rank	Sum of ranks
Ratings in RYB.Emoji-Ratings in Baseline	Negative ranks	167[a]	191.02	31901.00
	Positive ranks	216[b]	192.75	41635.00
	Ties	672[c]		
	Total	1055		

a. Ratings in RYB.Emoji<Ratings in Baseline

b. Ratings in RYB.Emoji>Ratings in Baseline

c. Ratings in RYB.Emoji=Ratings in Baseline

The details of the ranked scores of collectivists in baseline and RYG.Star scales are shown in Table 5.7. The negative ranks indicate where the rating score in the baseline was higher than RYG.Star scale, the positive ranks indicate those ratings which are higher in RYG.Star scale than the baseline and the tied ranks indicate where the ratings did not differ. In 216 cases, the ratings were higher and in 149 cases, they were lower than the baseline scale. On the other hand, the rank statistics in Table 5.8 indicate that for 228 of the 1055 cases, collectivists' ratings in RYB.Star scale was higher whereas in 149 cases, they assigned lower ratings using RYB.Star scale. In the star-based scales, the majority of the ranks were tied and the mean and the sum of positive ranks are higher than that of negative ranks. Table 5.9 showed that, in 205 cases

the ranks were positive and in 160 cases the ranks were negative. According to Table 5.10, in 216 cases, the ratings in the experimental scale were higher than the baseline scale and in 167 cases, they were lower. In both the emoji scales, the majority of the ranks were tied. In RYG.Emoji scale, the mean of negative ranks is higher than positive rank but in RYB.Emoji scale, the mean of positive rank is higher than negative rank.

Table 5.11: 20 notable association rules for collectivists

No.	LHS	RHS	Support	Confidence	Lift
1	{Baseline = {4,5}, RYG.Emoji = {4,5}, RYB.Emoji = {4,5}}	{RYB.Star={4,5}}	0.42	0.98	1.72
2	{Baseline = {3}, RYG.Star = {3}, RYB.Emoji = {3}, RYB.Star = {3}}	{RYG.Emoji={3}}	0.10	0.97	3.64
3	{Baseline = {4,5}, RYG.Star = {4,5}, RYG.Emoji = {4,5}}	{RYB.Emoji={4,5}}	0.42	0.97	1.74
4	{Baseline = {4,5}, RYG.Star = {4,5}, RYB.Emoji = {4,5}}	{RYG.Emoji={4,5}}	0.42	0.97	1.77
5	{Baseline = {4,5}, RYG.Emoji = {4,5}}	{RYG.Star={4,5}}	0.43	0.97	1.68
6	{Baseline = {1,2}, RYG.Star = {1,2}, RYG.Emoji = {1,2}}	{RYB.Star={1,2}}	0.11	0.94	5.30
7	{Baseline = {1,2}, RYG.Star = {1,2}, RYB.Star = {1,2}}	{RYB.Emoji={1,2}}	0.11	0.93	4.71
8	{Baseline = {1,2}, RYB.Star = {1,2}}	{RYG.Star={1,2}}	0.12	0.91	4.96
9	{Baseline = {1,2}, RYB.Star = {1,2}}	{RYB.Emoji={1,2}}	0.12	0.91	4.60
10	{Baseline = {3}, RYG.Star = {4,5}}	{RYB.Star = {4,5}}	0.07	0.84	1.46
11	{Baseline = {3}, RYG.Star = {4,5}, RYB.Star = {4,5}}	{RYB.Emoji = {4,5}}	0.05	0.84	1.50
12	{Baseline = {3}, RYB.Star = {4,5}}	{RYG.Star = {4,5}}	0.07	0.83	1.44
13	{Baseline = {3}, RYB.Emoji = {4,5}}	{RYG.Star = {4,5}}	0.06	0.83	1.44
14	{Baseline = {3}, RYG.Emoji = {4,5}}	{RYB.Emoji = {4,5}}	0.05	0.82	1.47
15	{Baseline = {3}, RYG.Emoji = {4,5}}	{RYB.Star = {4,5}}	0.05	0.80	1.40
16	{Baseline = {3}, RYG.Star = {4,5}, RYB.Emoji = {4,5}}	{RYG.Emoji = {4,5}}	0.05	0.79	1.45
17	{Baseline = {3}, RYB.Emoji = {4,5}}	{RYB.Star = {4,5}}	0.06	0.79	1.38
18	{Baseline = {3}, RYG.Emoji = {4,5}}	{RYG.Star = {4,5}}	0.05	0.79	1.37
19	{Baseline = {3}, RYG.Star = {4,5}}	{RYB.Emoji = {4,5}}	0.06	0.77	1.39
20	{RYG.Emoji = {1,2}, RYB.Emoji = {1,2}}	{Baseline = {1,2}}	0.12	0.77	3.62

In order to gain an understanding of the score being adjusted and discover the specific direction of the score adjustment resulting from the impact of color-coded scales on collectivists, I employed the Apriori algorithm. It produced 91 frequent itemsets and 214 association rules, all satisfying the minimum requirements for support, confidence and lift. The value of minimum support was set to 0.05, minimum confidence to 0.75 and minimum lift to 1.0. Since the size of the database for each of the two groups is larger than the personality groups, I lowered the value of minimum support from 0.1 to 0.05 to ensure that enough itemsets are found. The association rules representing the rating behavior of collectivist consumers are shown in Table 5.11.

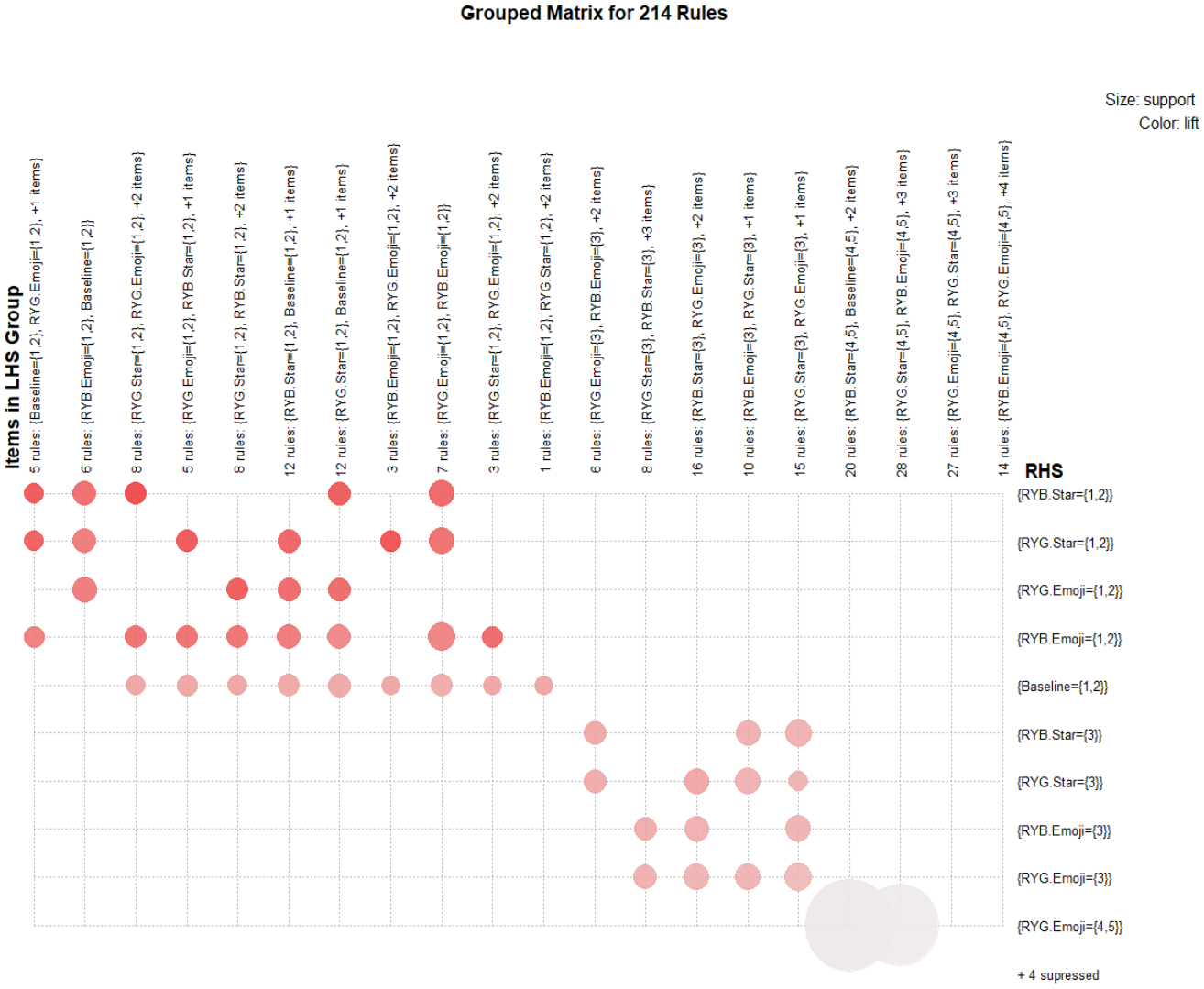


Figure 5.7: Influence of color-coded rating scales on collectivists (rules with confidence ≥ 0.75)

The rules in Table 5.11 provided the direction of the participants' score adjustment. As stated by rule 10, if a collectivist rated neutrally in baseline and gave a high rating score in RYG.Star scale, it infers that he will provide a high score in RYB.Star scale. Rule 12 also captures a similar pattern, if a consumer has

given a neutral rating in the baseline scale and a high rating in RYB.Star scale, it implies that he will rate the same product high using RYG.Star scale. Rules 10 to 19 clearly reflect on the existence of rating bias among collectivists since they evidently adjusted their true neutral ratings towards the high endpoints in all four experimental scales. Although the rating bias in collectivists' rating pattern is not as prominent as extroverts' since the minimum support threshold is higher in the latter's. However, the genuine high or low ratingss did not get impacted by the color-coded rating scales. For example, according to rule 3, if a consumer has given a high score in baseline and also given a high score in RYG.Star and RYG.Emoji scales, it also implies that he will rate the product similarly in RYB.Emoji scale. According to rule 6, the true low rating given in the baseline did not differ across RYG.Star, RYG.Emoji and RYB.Star scales. Figure 5.7 presents the grouped matrix-based visualization of the association rules for collectivists. The antecedents (LHS) of the 214 rules are plotted along the horizontal axis and the consequents (RHS) are plotted along the vertical axis. The size and the color of the bubble indicate the support and the lift respectively. The most interesting groups are placed in the top-left corner of the plot, for example, one of them consists of 5 rules which contain "Baseline={1,2}," and "RYG.Emoji={1,2}" and an additional item in the antecedent and the consequent is "RYB.Star={1,2}".

In summary, the color-coded neutral scales (RYG.Star scale and RYB.Star scale) influenced the rating behavior of extroverts, while no other personality traits, in the user study, experienced such an effect. On the other hand, collectivists showed biased rating patterns to a moderate extent in the presence of color-coded rating scales (RYG.Star scale, RYB.Star scale, RYG.Emoji scale and RYB.Emoji scale) but individualists exhibited no such biased rating behavior.

5.4 Summary of the Analysis

The key findings of the analysis are summarized below:

1. The differences between the ratings given by extroverts in the baseline scale and RYG.Star scale are statistically significant. It was also inferred from the inferential statistics that under the influence of RYG.Star scale, extroverts tend to assign higher scores than their true ratings.

2. For extroverts, the differences between the ratings in the baseline scale and RYB.Star scale are also statistically significant. The inferential statistics also demonstrated that the influence exerted by the RYB.Star scale would manipulate extroverts to assign higher scores than their original evaluation scores. Interestingly, both the emoji-based scales failed to influence their true ratings.

3. The association rules revealed that, due to the bias, impartial raters would get encouraged to provide an extreme response and consequently would shift their original score to the higher endpoints of the star-based scales shaded with green or blue color. Surprisingly, the original extreme responses given in the baseline scale were not affected by the bias.

4. Collectivists exhibited rating bias in the presence of the color effect. The ratings assigned by them in the four experimental scales were statistically significantly different from their respective ratings in the baseline scale. Eventually, the rating bias persuaded the respondents to assign comparatively higher scores in RYG.Star, RYB.Star, RYG.Emoji and RYB.Emoji scales than they would originally assign using the baseline scale.

5. The association rules revealed that, due to the color effect, the real impartial raters would adjust their original score towards the higher endpoints of the star or emoji-based scales shaded with green or blue color. However, the extreme responses (a low or a high rating score) given in the baseline scale were not affected by the bias.

5.5 Discussion

This section discusses the key findings and their implications in the analysis of rating bias induced by color-coded rating scales. On the basis of these findings, this section also answers the four research questions of this thesis.

5.5.1 Do consumers with different personality traits utilize similar color-coded rating scale differently for the same product?

The analyses in this thesis proved that indeed, consumers' utilization style for a color-based rating scale would be different based on their personality traits. Simple visual inspection of the descriptive statistics summarized by Figure 5.2 and 5.3 provides an overall understanding of the measure of central tendency and the measure of variability of the ratings provided by different personality groups. But the question remains, whether individuals' choices of feedback are heterogeneous across different color-coded rating scales for the same product. To find out, if there is any significant consensus between the ratings given in the baseline and the experimental scales, the non-parametric Wilcoxon signed rank test was employed. Although the initial intuition perceived from the literature gives the notion of potential rating bias to some extent among participants of the personality groups of extraversion, introversion, agreeableness and emotional stability. Interestingly, none of the statistical difference between the ratings was confirmed as significant by the non-parametric Wilcoxon signed rank test, except for the rating behavior of extroverts. It was apparent from the test that, extroverts exhibited biased rating behavior in both the neutral scales with endpoints shaded with two different colors. In fact, in more than 50% of the cases, they rated a product higher in RYG.Star scale and RYB.Star scale than they would in the most commonly used monochromatic star scale. This confirms the findings of [75], but only with respect to extroverts. Since extroverts are more assertive and receptive to new experience, they were more susceptible to the influence of the presence of contrasting color schemes at two endpoints of the scales and therefore, shifted their score to the higher endpoints of the scales. It implies that for a system that collects user ratings by means of a color-coded star-based scale, it will be difficult to

know the genuine opinion of extroverts. On the contrary, scales with expressive icons or the "human" scales, such as RYG.Emoji and RYB.Emoji scales failed to manipulate their true ratings. In my opinion, the bias was only significant in the "neutral" scales because of their familiarity, which possibly stimulated spontaneity among the participants and helped them stay less mindful while providing their feedback. Since expressive or descriptive icons are less commonly used in rating scales than stars, they clearly induced pondered ratings and therefore, participants might have been comparatively more mindful while utilizing emoji-based scales. No other speculation on the rating bias resulting from personality-wise color preference was confirmed by the analysis. This is probably because prior findings on personality-wise color preference in a general context might not necessarily reflect users' color preference in the context of UI design.

Since the overall evaluation concludes that the contrasting color schemes in the "neutral" scales or the star-based scales are responsible for impacting the rating behavior of extroverts only, whereas they did not play any role in distorting genuine feedback of participants with other personality traits. Therefore, it implies that consumers with different personality traits will utilize a "neutral" scale that has endpoints shaded with contrasting colors differently for the same product. These findings lead researchers to question the validity of ratings given by individuals using a one-size-fits-all rating scale and draw attention to the possibility of the system to fall short of reflecting on the genuine feedback of users with different personality traits.

5.5.2 Do collectivist consumers utilize similar color-coded rating scale differently from individualist consumers?

This research provided evidences that, while utilizing a color-coded scale, collectivists' rating behavior would be different from individualists. The non-parametric Wilcoxon signed rank test confirmed that, with regards to collectivists, the ratings given to the same item in different color-coded rating scales are statistically significantly different from that in the baseline. The impact of colors used in both the "neutral" and "human" scales is responsible for collectivists' biased rating behavior. Although the distortion in ratings resulting from the influence of the color-coded scales was statistically significant and biased to the higher endpoints of the scales, interestingly, in more than 50% of the cases the influence was ineffective and the ratings in the color-coded scales were indifferent from the baseline. In accordance with the rationale deduced from the related works discussed in Chapter 3, the result proved that due to collectivists' preference for colorful and appealing interface, the colors at the endpoints exerted a cognitive influence on the numeric interpretation of the scale. However, the influence did not persuade them to compress the distribution of the ratings towards the central area of the scale. In agreement with the decisions in [75], the ratings were shifted to the higher endpoints of the scales. On the contrary, according to the study, the same characteristics did not exert any influence on individualists. This is probably because of the lack of preference in individualists for a colorful and visually appealing interface.

The overall evaluation corroborated the claim of this research that collectivists' predilection for colorful and visually appealing decision-making interface influenced their rating decisions and hence they provided

biased rating scores in RYG.Star, RYB.Star, RYG.Emoji and RYB.Emoji scales. But interestingly, individualists are resistant to such influence of the scales. This observation answers the research question *"Do collectivist consumers utilize similar color-coded rating scale differently from individualist culture?"* and establishes that collectivists rate a certain product differently in a color-coded rating scale than individualists.

5.5.3 In case of a biased rating, how do consumers adjust their actual ratings?

In summary, the association rules provided a precise direction of how extroverts and collectivists would modify their true ratings scores under the influence of colors in rating scales. To some extent, the results of this research resonated with the findings of [75] and disapproved the claims in [13].

The analysis done so far established extroverts' and collectivists' proneness to the impact of colors in rating scales. Yet it did not provide any specification about the nature of the score being modified and any pinpoint direction of the score adjustment resulting from the impact. So, to discover more specific details, the Apriori algorithm was applied to the ratings provided by the extroverts and the collectivists. The interesting patterns produced by the algorithm revealed that, under the influence of the contrasting colors of the rating scales, extroverts changed their impartial responses to high scores. To explain, they shifted their genuine neutral score to the higher endpoints of the RYG.Star and RYB.Star scales. Likewise, in RYG.Star, RYB.Star, RYG.Emoji and RYB.Emoji scales, collectivists shifted from a score which was originally supposed to be a neutral score to a high rating score. Interestingly, in both the groups, the score adjustment of an individual was directed in the same direction. The association rules also disclosed that a genuine low or high score in the baseline would not be manipulated by the influence of the color-coded rating scales.

5.5.4 Can a personality and culture-based approach clarify the contradictory rating behaviors observed in the literature review?

Yes, a personality and culture-based approach clarified the contradictory patterns in consumers' rating behavior to some extent. As summarized in Table 3.2, the results of the existing research works investigating the impact of colors used in a rating scale did not provide an elaborate and unanimous decision regarding the bias. After conducting a survey, Tourangeau et al. [75] concluded that due to bias, users shifted their rating scores towards the upper bound of the scale. On the contrary, Bonaretti et al. [13] stated that the influence exerted by the color-coded rating scales would most likely impel users to adjust their ratings to the central point of the scale. The exploratory study in my research clarified such contradiction by taking a personality and culture-based approach to analyzing users' rating behavior. It found that the influence of colors of a rating scale is not uniformly strong for every reviewer. For example, star-based scales built with contrasting colors were highly influential on extroverts while for participants with other personality traits, the scales were impactless. Interestingly, participants from each of the personality groups reacted similarly to both the emoji-based scales. The impact of the rating scales was also perceived differently by collectivists from individualists. While collectivist participants exhibited biased rating behavior under the influence of

the color-coded “neutral” and “human” scales, individualist participants were completely invulnerable to the influence.

Unlike earlier works, the analysis of this research is elaborate; it provides a thorough understanding of the score being subject to the adjustment and the precise direction of the rating score adjustment resulting from the bias. While existing works depicted a generalized picture of score alteration, this research provides evidence that score alteration is not applicable for an extreme response and only directed from the neutral to the higher end of the scale. This research ascribes the contradictory patterns to the variance in personality and culture-based interpretation and the results corroborated this claim. Therefore, it can be inferred that users’ approaches to utilizing a color-coded rating scale are diverse and can be mapped to their personality and culture.

6 Conclusion and Future work

Post-consumption ratings provided by consumers have become an integral part of the data-driven systems. Ratings provide information about consumers' overall satisfaction with the products and assist the system to tailor contents in accordance with users' preferences. Hence, both the consumers and the system rely on ratings to a great extent. The efficiency of the system depends on the quality and originality of user ratings. Earlier research works have explored the impact of rating scale characteristics (e.g. neutral point, labeling, color) on users' responses. For example, labeling all the points of a rating scale can lower respondents' tendency of providing extreme responses. In the course of investigating the impact of color of rating scales on users' ratings, several researchers have discovered contradictory patterns of bias in their rating behavior. The lack of a unanimous decision on the rating bias in color-coded scales highlights the necessity of taking users' individuality into account. This research argued the unsuitability of the general approach since the cognitive impact on users' interpretation of a scale might differ based on their color preferences and susceptibility to the impact. Moreover, the existing works did not investigate the sole impact of colors on users' rating behavior, instead they amalgamated other characteristics of a scale with colors.

This research work investigated whether the presence of colors in rating scales can impact users' responses differently depending on their personality and culture. To this aim, I designed a user study integrated with the Big Five survey of 44 items, a demographic survey and a rating collection process. The rating collection was conducted using the commonly seen yellow-star scale as the baseline and four color-coded experimental scales. Raters provided their rating scores using these scales only for the products they have consumed or experienced previously. The collected data about users' personality, culture and post-consumption ratings were then analyzed to observe their personality and culture-wise rating patterns across different color-coded scales.

The analysis showed that using color-coded star scales for collecting user ratings can significantly impact the original evaluation of extroverts and persuade them to provide a higher score. Because of extroverts' assertive and enthusiastic nature, the scales' cognitive impact most likely manipulated their numerical interpretation of the scale and convinced them to change their "true" ratings. Collectivists, however, exhibited less intensive biased ratings than extroverts in all four experimental scales. This infers that their preference for colorful and appealing interface clouded their judgment of the numerical interpretation of the scales and eventually caused the biased behavior. Furthermore, the pattern mining technique that was applied to the collected ratings shed light on how individuals altered their true ratings because of the influence of color-coded rating scales.

The findings of this research provided important insights on the diversity in the rating scale utilization style of users with different personalities and cultures. It proved that the color of a scale can solely be responsible for instigating bias in users' ratings which ultimately cast doubt on the authenticity of the ratings given by consumers in a system that uses a color-coded one-size-fits-all rating scale to collect user feedback. This research also showed strong associations among users' rating behaviors across different color-based scales which can potentially contribute to the mechanism of converting rating scores across different platforms. Finally, on the ground of the key findings from the analysis, this research also suggested design guidelines for different data-driven systems.

6.1 Design Recommendations

Taking notes from the insights into the rating scale utilization style of users with different personality and culture, I offer the following design recommendations for a data-driven system:

Identifying the validity of ratings is important considering the compromised ratings might expose the system to the risk of inefficiency and failure to infer the actual interests of the consumers. When associated with users' individuality, their ratings can provide in-depth information about how valuable and genuine the feedbacks actually are. On the grounds of the findings, this research suggests that when intelligent systems such as e-commerce websites, recommender systems, online communities collect user feedback, it is critical to take users' individuality into account. In order to collect more valid and useful feedback, designers can employ mechanisms to implicitly acquire an individual's personality and culture and consider taking a personalized, targeted approach to designing the decision environment by leveraging that information. For example, extroverts and collectivists are likely to give high rating scores in the red-yellow-green star scale. The graphical interface of the system adopting that scale to collect ratings of the consumers could attempt to make extroverts and collectivists more cognizant of the potential rating biases and motivate them to give a true evaluation. Upon identifying extroverts and collectivists, the system can also offer a personalized rating scale which would not instigate any bias, instead of a scale targeting the general population. The insights can also contribute to the mitigation of bias. Existing research works proposed methods to correct the bias by rescaling the compromised rating scores [61, 30], but the proposed methods reflected the general population and ignored the aspect of user's individuality. Taking the personality and culture-based approach to mitigate the bias hidden in the rating scores can narrow down the intricate process of bias mitigation. Therefore, it is very important to take these factors into consideration for building a data-driven system.

6.2 Limitations and Future Works

Besides the contributions made by the thesis, this work has some potential limitations. In terms of the study procedure, this work is limited by the fact that it was conducted with a small sample size of respondents. As a consequence, the small-sized groups of participants who are emotionally stable or introvert might have led

us to a biased conclusion about their rating behavior. Moreover, none of the subjects was classified into the groups with high closeness to experience, antagonism and lack of direction, therefore, a comparison within each of these three groups could not be conducted. Another limitation of the thesis is that the dataset is not prepared to fit the algorithm for capturing the smallest changes in case of a score alteration. For instance, a score adjustment from 4 to 5 points is considered as an altered score in the Wilcoxon signed rank test, whereas, in the rule mining technique, this change is still considered indifferent and under the same group representing a high score. Furthermore, individuals' culture might evolve based on the society they have been living in for the longest period of time. For instance, an individual might be born and brought up in a collectivist society, but they might adopt to individualism considering that they are continually residing in a country which follows an individualistic culture. However, considering the age range of the majority of the participants of this study, this constraint did not have any significant impact on the findings of the study.

Based on the limitations, we proposed a number of scopes for future extensions. A follow-up study can be conducted in the future with a larger group of respondents with more diversified personality traits, to capture potential biases hidden in their ratings. There is also scope for extending the study in the future to observe the difference between users' responses in emoji scales and star scales with monochromatic color schemes. Moreover, the dataset has further scope to be redesigned in a way so that it can capture the smallest score deviation with various degrees of alteration in the rating score. In future, the demographic questionnaire can be redesigned to consider the cases of participants' cultures which they are either born into or adapted to. Future work could also expand by investigating the effectiveness of the strategies to educate the users of a system about the potential biases with the help of graphical interfaces and the effectiveness of the proposed design implications for a personalized rating interface with the capability of recognizing and avoiding potential rating bias.

References

[1] Hofstede's cultural framework. https://opentextbc.ca/principlesofmanagementopenstax/chapter/hofstedes-cultural-framework/. [Online; accessed 11-December-2020].

[2] The emotional value of color. http://www.tek-unique.com/the-emotional-value-of-color/, 2009. [Online; accessed 20-May-2020].

[3] Ifeoma Adaji, Kiemute Oyibo, and Julita Vassileva. Understanding low review ratings in online communities: A personality based approach. In *PPT@ PERSUASIVE*, pages 34–42, 2018.

[4] Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjng Zhang. Reducing recommender systems biases: An investigation of rating display designs. *Forthcoming, MIS Quarterly*, pages 19–18, 2019.

[5] Gediminas Adomavicius, Jesse C Bockstedt, Shawn P Curley, and Jingjing Zhang. Do recommender systems manipulate consumer preferences? a study of anchoring effects. *Information Systems Research*, 24(4):956–975, 2013.

[6] Agnes Ogee. How to correctly interpret p values. https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-correctly-interpret-p-values, 2014. [Online; accessed 19-July-2020].

[7] Fawzi Abdulkhaliq Alghamdi, Amina Tariq, Talal Naveed Puri, et al. Relative ranking–a biased rating. In *Innovations and Advances in Computer Sciences and Engineering*, pages 25–29. Springer, 2010.

[8] American Psychological Association. Personality. https://www.apa.org/topics/personality/, 2019. [Online; accessed 22-May-2020].

[9] Taiwo Amoo and Hershey H Friedman. Do numeric values influence subjects' responses to rating scales? *Journal of International Marketing and Marketing Research*, 26:41–46, 2001.

[10] Andrew McCaskill. Recommendations from friends remain most credible form of advertising among consumers; branded websites are the second-highest-rated form. https://www.nielsen.com/eu/en/press-releases/2015/recommendations-from-friends-remain-most-credible-form-of-advertising/, 2015. [Online; accessed 20-May-2020].

[11] Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. Personality and patterns of facebook usage. In *Proceedings of the 4th annual ACM web science conference*, pages 24–32, 2012.

[12] Gary E Bolton, Elena Katok, and Axel Ockenfels. How effective are electronic reputation mechanisms? an experimental investigation. *Management science*, 50(11):1587–1602, 2004.

[13] Dario Bonaretti, Marcin Lukasz Bartosiak, and Gabriele Piccoli. Cognitive anchoring of color cues on online review ratings. In *23rd Americas Conference on Information Systems, AMCIS 2017, Boston, MA, USA, August 10-12, 2017*. Association for Information Systems, 2017.

[14] Briana Brownell. Do colours affect survey responses? https://insightrix.com/colours-affect-survey-responses/, 2019. [Online; accessed 21-May-2020].

[15] Federica Cena, Cristina Gena, Pierluigi Grillo, Tsvi Kuflik, Fabiana Vernero, and Alan J Wecker. How scales influence user rating behaviour in recommender systems. *Behaviour & Information Technology*, 36(10):985–1004, 2017.

[16] Federica Cena and Fabiana Vernero. A study on user preferential choices about rating scales. *International Journal of Technology and Human Interaction (IJTHI)*, 11(1):33–54, 2015.

[17] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, 17(3):433–450, 2013.

[18] Dan Cosley, Shyong K Lam, Istvan Albert, Joseph A Konstan, and John Riedl. Is seeing believing? how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 585–592, 2003.

[19] Daniella Alscher. Color contrast: For the sake of aesthetic and accessibility. https://learn.g2.com/color-contrast, 2019. [Online; accessed 05-July-2020].

[20] Jorge Carrillo De Albornoz, Laura Plaza, Pablo Gervás, and Alberto Díaz. A joint model of feature mining and sentiment analysis for product review rating. In *European conference on information retrieval*, pages 55–66. Springer, 2011.

[21] Mehmet Deniz. An investigation of decision making styles and the five-factor personality traits with respect to attachment styles. *Educational Sciences: Theory and Practice*, 11(1):105–113, 2011.

[22] Elise Moreau. The top social networking sites people are using. https://www.lifewire.com/top-social-networking-sites-people-are-using-3486554, 2020. [Online; accessed 19-July-2020].

[23] Ron Garland. The mid-point on a rating scale: Is it desirable. *Marketing bulletin*, 2(1):66–70, 1991.

[24] Diana Gavilan, Maria Avello, and Gema Martinez-Navarro. The influence of online ratings and reviews on hotel booking consideration. *Tourism Management*, 66:53–61, 2018.

[25] Geert Hofstede. Hofstede insights. https://www.hofstede-insights.com/country-comparison/. [Online; accessed 19-July-2020].

[26] Cristina Gena, Roberto Brogi, Federica Cena, and Fabiana Vernero. The impact of rating scales on user's rating behavior. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 123–134. Springer, 2011.

[27] David Godes and Dina Mayzlin. Using online conversations to study word-of-mouth communication. *Marketing science*, 23(4):545–560, 2004.

[28] Jennifer Golbeck and Eric Norris. Personality, movie preferences, and recommendations. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1414–1415, 2013.

[29] Carlos A Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4):1–19, 2015.

[30] Eric A Greenleaf. Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, 29(2):176–188, 1992.

[31] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[32] Geert Hofstede. The cultural relativity of organizational practices and theories. *Journal of international business studies*, 14(2):75–89, 1983.

[33] Geert Hofstede. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications, 2001.

[34] Joshua N Hook, Everett L Worthington Jr, and Shawn O Utsey. Collectivism, forgiveness, and social harmony. *The Counseling Psychologist*, 37(6):821–847, 2009.

[35] Rong Hu and Pearl Pu. Exploring relations between personality and user rating behaviors. In *UMAP Workshops*, 2013.

[36] Jacci Howard Bear. Learn the basics of contrasting colors on the color wheel. https://www.lifewire.com/contrasting-colors-in-design-1078274, 2019. [Online; accessed 05-July-2020].

[37] Gawesh Jawaheer, Martin Szomszor, and Patty Kostkova. Comparison of implicit and explicit feedback from an online music recommendation service. In *proceedings of the 1st international workshop on information heterogeneity and fusion in recommender systems*, pages 47–51, 2010.

[38] Lucas J Jenkins, Yung-Jui Yang, Joshua Goh, Ying-Yi Hong, and Denise C Park. Cultural differences in the lateral occipital complex while viewing incongruent scenes. *Social cognitive and affective neuroscience*, 5(2-3):236–241, 2010.

[39] Joe Hopper. Don't color-code your nps net promoter scale. https://verstaresearch.com/blog/dont-color-code-your-nps-net-promoter-scale/, 2019. [Online; accessed 21-May-2020].

[40] Oliver P John, Eileen M Donahue, and Robert L Kentle. Big five inventory. *Journal of Personality and Social Psychology*, 1991.

[41] Oliver P John, Sanjay Srivastava, et al. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999.

[42] Katarzyna Kabzińska, Magdalena Wieloch, Dominik Filipiak, and Agata Filipowska. Profiling user's personality using colours: connecting bfi-44 personality traits and plutchik's wheel of emotions. In *International Conference on Information Systems Architecture and Technology*, pages 371–380. Springer, 2018.

[43] Naveen K Kambham, Kevin G Stanley, and Scott Bell. Predicting personality traits using smartphone sensor data and app usage data. In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 125–132. IEEE, 2018.

[44] Raghav Pavan Karumur, Tien T Nguyen, and Joseph A Konstan. Exploring the value of personality in predicting rating behaviors: a study of category preferences on movielens. In *Proceedings of the 10th ACM conference on recommender systems*, pages 139–142, 2016.

[45] Raghav Pavan Karumur, Tien T Nguyen, and Joseph A Konstan. Personality, user preferences and behavior in recommender systems. *Information Systems Frontiers*, 20(6):1241–1265, 2018.

[46] Gurneet Kaur and Neena Madan. Association rule mining: A survey. *International Journal of Computer Science and Information Technologies*, 5(2):2320–2324, 2014.

[47] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[48] Michal Kosinski, Yoram Bachrach, Pushmeet Kohli, David Stillwell, and Thore Graepel. Manifestations of user personality in website choice and behaviour on online social networks. *Machine learning*, 95(3):357–380, 2014.

[49] Tsvi Kuflik, Alan J Wecker, Federica Cena, and Cristina Gena. Evaluating rating scales personality. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 310–315. Springer, 2012.

[50] Sarah Küsgen and Sören Köcher. The influence of customer product ratings on purchase decisions: An abstract. In *Creating Marketing Magic and Innovative Future Marketing Trends*, pages 953–954. Springer, 2017.

[51] Michel Laroche, Maria Kalamas, and Mark Cleveland. " i" versus" we": How individualists and collectivists use information sources to formulate their service expectations. *International marketing review*, 22(3):279–308, 2005.

[52] Sandra Larrivee, Frank L Greenway, and William D Johnson. A statistical analysis of a traffic-light food rating system to promote healthy nutrition and body weight. *Journal of diabetes science and technology*, 9(6):1336–1341, 2015.

[53] Hock-Eam Lim. The use of different happiness rating scales: Bias and comparison problem? *Social Indicators Research*, 87(2):259–267, 2008.

[54] Gitte Lindgaard, Cathy Dudek, and Gerry Chan. Cultural congruence and rating scale biases in home-pages. In *IFIP Conference on Human-Computer Interaction*, pages 531–538. Springer, 2013.

[55] Richard Lowry. Concepts and applications of inferential statistics. 2014.

[56] Michael Luca. Reviews, reputation, and revenue: The case of yelp. com. *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, (12-016), 2016.

[57] Warih Maharani, Dwi H Widyantoro, and Masayu L Khodra. Discovering users' perceptions on rating visualizations. In *Proceedings of the 2nd International Conference in HCI and UX Indonesia 2016*, pages 31–38, 2016.

[58] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

[59] Susan M Mudambi, David Schuff, and Zhewei Zhang. Why aren't the stars aligned? an analysis of online review content and star ratings. In *2014 47th Hawaii International Conference on System Sciences*, pages 3139–3147. IEEE, 2014.

[60] Kiemute Oyibo, Yusuf Sahabi Ali, and Julita Vassileva. An empirical analysis of the perception of mobile website interfaces and the influence of culture. In *PPT@ PERSUASIVE*, pages 44–56, 2016.

[61] Kyung-Wha Park, Byoung-Hee Kim, Tae-Suh Park, and Byoung-Tak Zhang. Uncovering response biases in recommendation. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*. Citeseer, 2014.

[62] Pon Staff. The anchoring effect and how it can impact your negotiation. https://www.pon.harvard.edu/daily/negotiation-skills-daily/the-drawbacks-of-goals/, 2019. [Online; accessed 20-May-2020].

[63] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 180–185. IEEE, 2011.

[64] Daniele Quercia, Renaud Lambiotte, David Stillwell, Michal Kosinski, and Jon Crowcroft. The personality of popular facebook users. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 955–964, 2012.

[65] Raluca Budiu. Between-subjects vs. within-subjects study design. https://www.nngroup.com/articles/between-within-subjects/, May 13, 2018. [Online; accessed 11-December-2020].

[66] Rashmi Jain. A beginner's tutorial on the apriori algorithm in data mining with r implementation. https://www.hackerearth.com/blog/developers/beginners-tutorial-apriori-algorithm-data-mining-r-implementation/, 2017. [Online; accessed 19-July-2020].

[67] Michel Raymond and François Rousset. An exact test for population differentiation. *Evolution*, 49(6):1280–1283, 1995.

[68] Amit Sharma, Jake M Hofman, and Duncan J Watts. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 453–470, 2015.

[69] Jianqiang Shen, Oliver Brdiczka, and Juan Liu. Understanding email writers: Personality prediction from email messages. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 318–330. Springer, 2013.

[70] Richard M Steers. *Introduction to organizational behavior*. Goodyear Publishing Company, 1981.

[71] Stephanie Glen. Order effects: Definition, examples and solutions. https://www.statisticshowto.com/order-effects/, 2019. [Online; accessed 19-July-2020].

[72] Stephanie Glen. Wilcoxon signed rank test: Definition, how to run. https://www.statisticshowto.com/wilcoxon-signed-rank-test/, 2020. [Online; accessed 19-July-2020].

[73] Huatong Sun. Building a culturally-competent corporate web site: an exploratory study of cultural markers in multilingual web design. In *Proceedings of the 19th annual international conference on Computer documentation*, pages 95–102, 2001.

[74] Teo Choong Ching. Types of cognitive biases you need to be aware of as a researcher. https://uxdesign.cc/cognitive-biases-you-need-to-be-familiar-with-as-a-researcher-c482c9ee1d49, 2016. [Online; accessed 20-May-2020].

[75] Roger Tourangeau, Mick P Couper, and Frederick Conrad. Color, labels, and interpretive heuristics for response scales. *Public Opinion Quarterly*, 71(1):91–112, 2007.

[76] Sonja Treven, Matjaž Mulej, and Monty Lynn. The impact of culture on organizational behavior. *Management: journal of contemporary management issues*, 13(2 (Special issue)):27–39, 2008.

[77] Dimitrios Tsekouras. The effect of rating scale design on extreme response tendency in consumer product ratings. *International Journal of Electronic Commerce*, 21(2):270–296, 2017.

[78] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131, 1974.

[79] Jeroen Van Barneveld and Mark Van Setten. Designing usable interfaces for tv recommender systems. In *Personalized Digital Television*, pages 259–285. Springer, 2004.

[80] Paula Cristina Vaz, Ricardo Ribeiro, and David Martins De Matos. Understanding the temporal dynamics of recommendations across different rating scales. In *UMAP Workshops*, 2013.

[81] Bert Weijters, Elke Cabooter, and Niels Schillewaert. The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27(3):236–247, 2010.

[82] Magdalena Wieloch, Katarzyna Kabzińska, Dominik Filipiak, and Agata Filipowska. Profiling user colour preferences with bfi-44 personality traits. In *International Conference on Business Information Systems*, pages 63–76. Springer, 2018.

[83] Yonnie Chyung, Ieva Swanson. Evidence-based survey design: The use of sliders. https://www.td.org/insights/evidence-based-survey-design-the-use-of-sliders, 2019. [Online; accessed 1-July-2020].

[84] Jingjing Zhang. Anchoring effects of recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 375–378, 2011.

Appendix A

Study Questionnaire

A.1 Demographic Questionnaire

Please provide the following information:

1. Age: [ ]
2. Please select your gender:
   - ☐ Male
   - ☐ Female
   - ☐ Other
3. Home Country: (The country where you were born and usually raised in, regardless of your present country of residence and citizenship.) [ ]

A.2 Big Five Questionnaire

1. Is talkative.
2. Tends to find fault with others.
3. Does a thorough job.
4. Is depressed, blue.
5. Is original, comes up with new ideas.
6. Is reserved.
7. Is helpful and unselfish with others.
8. Can be somewhat careless.
9. Is relaxed, handles stress well.
10. Is curious about many different things.
11. Is full of energy.
12. Starts quarrels with others.
13. Is a reliable worker.
14. Can be tense.
15. Is ingenious, a deep thinker.
16. Generates a lot of enthusiasm.
17. Has a forgiving nature.
18. Tends to be disorganized.
19. Worries a lot.
20. Has an active imagination.
21. Tends to be quiet.
22. Is generally trusting.
23. Tends to be lazy.
24. Is emotionally stable, not easily upset.
25. Is inventive.
26. Has an assertive personality.
27. Can be cold and aloof.
28. Perseveres until the task is finished.
29. Can be moody.
30. Values artistic, aesthetic experiences.
31. Is sometimes shy, inhibited.
32. Is considerate and kind to almost everyone.
33. Does things efficiently.
34. Remains calm in tense situations.
35. Prefers work that is routine.
36. Is outgoing, sociable.
37. Is sometimes rude to others.
38. Makes plans and follows through with them.
39. Gets nervous easily.
40. Likes to reflect, play with ideas.
41. Has few artistic interests.
42. Likes to cooperate with others.
43. Is easily distracted.
44. Is sophisticated in art, music, or literature.

Appendix B

User Interface of Big Five Survey

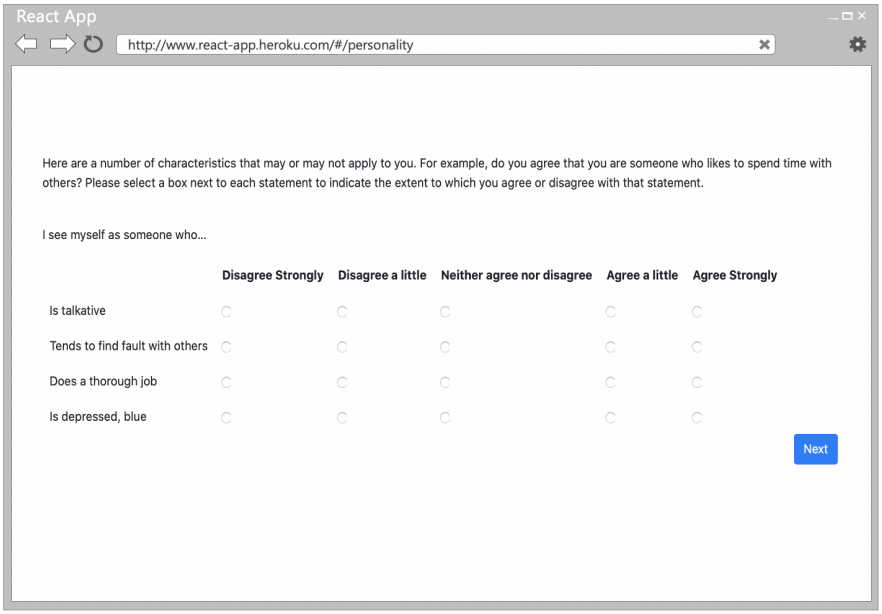


Figure B.1: User study interface for Big Five Survey: Page 1 (item 1-4)

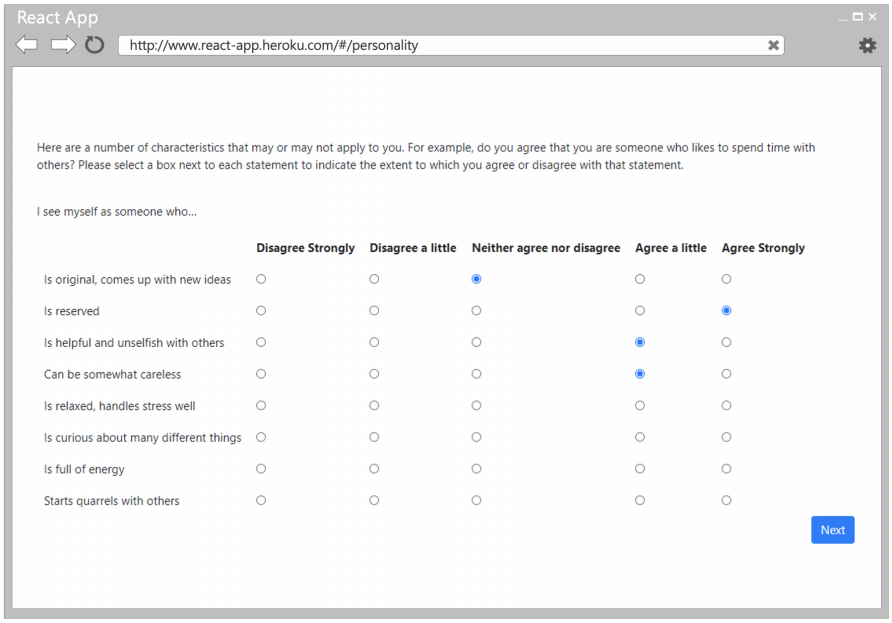


Figure B.2: User study interface for Big Five Survey: Page 2 (item 5-12)

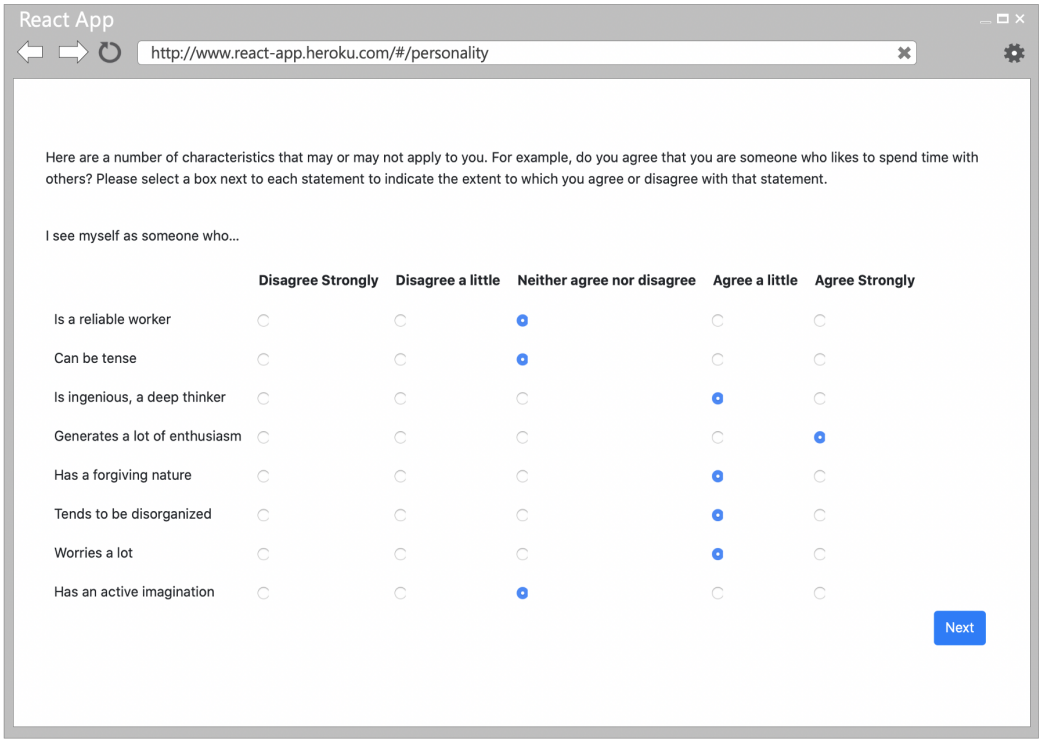


Figure B.3: User study interface for Big Five Survey: Page 3 (item 13-20)

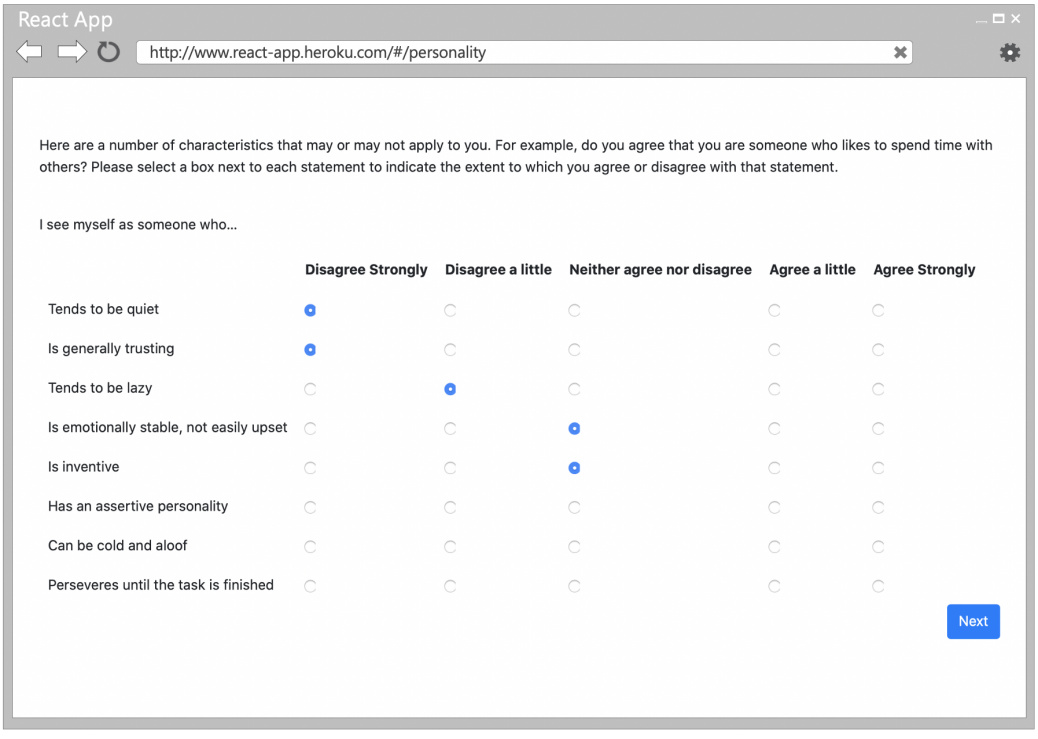


Figure B.4: User study interface for Big Five Survey: Page 4 (item 21-28)

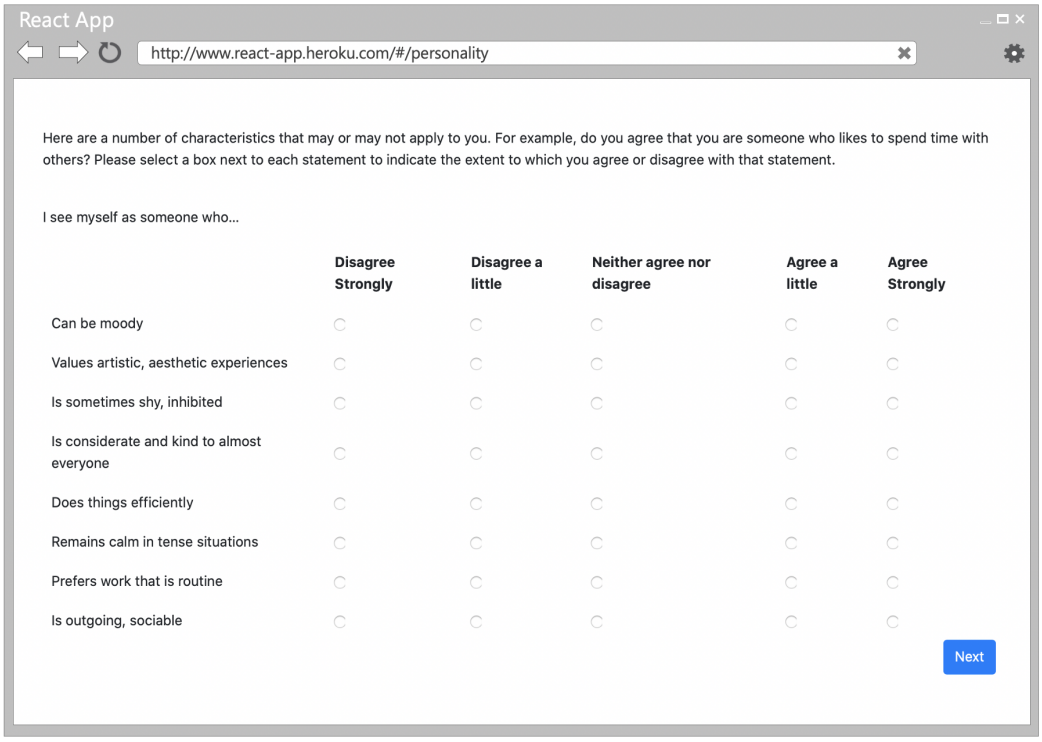


Figure B.5: User study interface for Big Five Survey: Page 5 (item 29-36)

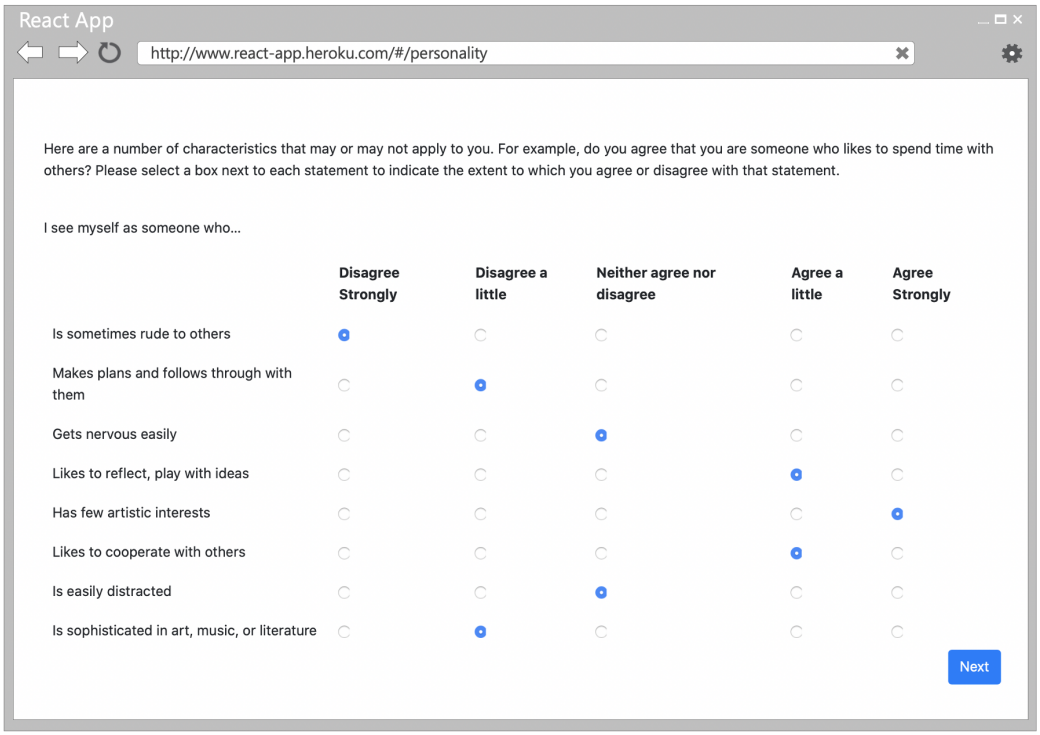


Figure B.6: User study interface for Big Five Survey: Page 6 (item 37-44)

Appendix C

Association Rules

C.1 142 Association Rules For Extroverts

Rules	Support	Confidence	Lift
{RYG.Star={1,2}} =>{RYB.Star={1,2}}	0.1525	0.7826	4.0151
{RYB.Star={1,2}} =>{RYG.Star={1,2}}	0.1525	0.7826	4.0151
{RYG.Star={1,2}} =>{RYG.Emoji={1,2}}	0.1525	0.7826	3.8478
{RYG.Emoji={1,2}} =>{RYG.Star={1,2}}	0.1525	0.75	3.8478
{RYB.Star={1,2}} =>{RYG.Emoji={1,2}}	0.1568	0.8043	3.9547
{RYG.Emoji={1,2}} =>{RYB.Star={1,2}}	0.1568	0.7708	3.9547
{RYB.Star={1,2}} =>{RYB.Emoji={1,2}}	0.1695	0.8696	3.6003
{RYG.Emoji={1,2}} =>{RYB.Emoji={1,2}}	0.1695	0.8333	3.4503
{RYG.Star={3}} =>{RYB.Star={3}}	0.1653	0.75	3.3396
{RYG.Star={3}} =>{RYG.Emoji={3}}	0.1695	0.7692	2.3887
{RYG.Star={3}} =>{RYB.Emoji={3}}	0.1653	0.75	2.2125
{RYB.Star={3}} =>{RYG.Emoji={3}}	0.1864	0.8302	2.578
{RYB.Star={3}} =>{RYB.Emoji={3}}	0.178	0.7925	2.3377
{RYG.Emoji={3}} =>{RYB.Emoji={3}}	0.2458	0.7632	2.2513
{RYB.Emoji={4,5}} =>{Baseline={4,5}}	0.3305	0.7879	1.8229
{Baseline={4,5}} =>{RYB.Emoji={4,5}}	0.3305	0.7647	1.8229
{RYB.Emoji={4,5}} =>{RYG.Emoji={4,5}}	0.3814	0.9091	1.9156
{RYG.Emoji={4,5}} =>{RYB.Emoji={4,5}}	0.3814	0.8036	1.9156
{RYB.Emoji={4,5}} =>{RYB.Star={4,5}}	0.3856	0.9192	1.5834
{RYB.Emoji={4,5}} =>{RYG.Star={4,5}}	0.3856	0.9192	1.572
{Baseline={4,5}} =>{RYB.Star={4,5}}	0.3305	0.7647	1.3173
{Baseline={4,5}} =>{RYG.Star={4,5}}	0.3432	0.7941	1.3581
{RYG.Emoji={4,5}} =>{RYB.Star={4,5}}	0.4364	0.9196	1.5842
{RYB.Star={4,5}} =>{RYG.Emoji={4,5}}	0.4364	0.7518	1.5842
{RYG.Emoji={4,5}} =>{RYG.Star={4,5}}	0.428	0.9018	1.5422
{RYB.Star={4,5}} =>{RYG.Star={4,5}}	0.5297	0.9124	1.5604
{RYG.Star={4,5}} =>{RYB.Star={4,5}}	0.5297	0.9058	1.5604
{RYG.Star={1,2},RYB.Star={1,2}} =>{RYG.Emoji={1,2}}	0.1398	0.9167	4.5069
{RYG.Star={1,2},RYG.Emoji={1,2}} =>{RYB.Star={1,2}}	0.1398	0.9167	4.7029
{RYG.Emoji={1,2},RYB.Star={1,2}} =>{RYG.Star={1,2}}	0.1398	0.8919	4.5758
{RYG.Star={1,2},RYB.Star={1,2}} =>{RYB.Emoji={1,2}}	0.1314	0.8611	3.5653
{RYG.Star={1,2},RYB.Emoji={1,2}} =>{RYB.Star={1,2}}	0.1314	0.9118	4.6777
{RYB.Emoji={1,2},RYB.Star={1,2}} =>{RYG.Star={1,2}}	0.1314	0.775	3.9761
{RYG.Star={1,2},RYB.Star={1,2}} =>{Baseline={1,2}}	0.1144	0.75	2.8548
{Baseline={1,2},RYG.Star={1,2}} =>{RYB.Star={1,2}}	0.1144	0.7941	4.0742
{Baseline={1,2},RYB.Star={1,2}} =>{RYG.Star={1,2}}	0.1144	0.7941	4.0742
{RYG.Star={1,2},RYG.Emoji={1,2}} =>{RYB.Emoji={1,2}}	0.1356	0.8889	3.6803
{RYG.Star={1,2},RYB.Emoji={1,2}} =>{RYG.Emoji={1,2}}	0.1356	0.9412	4.6275
{RYG.Emoji={1,2},RYB.Emoji={1,2}} =>{RYG.Star={1,2}}	0.1356	0.8	4.1043

Table C.1: Association rules for extroverts

Rules	Support	Confidence	Lift
{RYG.Star={1,2},RYG.Emoji={1,2}} =>{Baseline={1,2}}	0.1144	0.75	2.8548
{Baseline={1,2},RYG.Star={1,2}} =>{RYG.Emoji={1,2}}	0.1144	0.7941	3.9044
{Baseline={1,2},RYG.Emoji={1,2}} =>{RYG.Star={1,2}}	0.1144	0.7941	4.0742
{RYG.Star={1,2},RYB.Emoji={1,2}} =>{Baseline={1,2}}	0.1102	0.7647	2.9108
{Baseline={1,2},RYG.Star={1,2}} =>{RYB.Emoji={1,2}}	0.1102	0.7647	3.1662
{RYG.Emoji={1,2},RYB.Star={1,2}} =>{RYB.Emoji={1,2}}	0.1398	0.8919	3.6927
{RYB.Emoji={1,2},RYB.Star={1,2}} =>{RYG.Emoji={1,2}}	0.1398	0.825	4.0563
{RYG.Emoji={1,2},RYB.Emoji={1,2}} =>{RYB.Star={1,2}}	0.1398	0.825	4.2326
{Baseline={1,2},RYB.Star={1,2}} =>{RYG.Emoji={1,2}}	0.1144	0.7941	3.9044
{Baseline={1,2},RYG.Emoji={1,2}} =>{RYB.Star={1,2}}	0.1144	0.7941	4.0742
{Baseline={1,2},RYB.Star={1,2}} =>{RYB.Emoji={1,2}}	0.1229	0.8529	3.5315
{Baseline={1,2},RYB.Emoji={1,2}} =>{RYB.Star={1,2}}	0.1229	0.7632	3.9153
{RYG.Emoji={1,2},RYB.Emoji={1,2}} =>{Baseline={1,2}}	0.1271	0.75	2.8548
{Baseline={1,2},RYG.Emoji={1,2}} =>{RYB.Emoji={1,2}}	0.1271	0.8824	3.6533
{Baseline={1,2},RYB.Emoji={1,2}} =>{RYG.Emoji={1,2}}	0.1271	0.7895	3.8816
{RYG.Star={3},RYB.Star={3}} =>{RYG.Emoji={3}}	0.1525	0.9231	2.8664
{RYG.Star={3},RYG.Emoji={3}} =>{RYB.Star={3}}	0.1525	0.9	4.0075
{RYG.Emoji={3},RYB.Star={3}} =>{RYG.Star={3}}	0.1525	0.8182	3.7133
{RYG.Star={3},RYB.Star={3}} =>{RYB.Emoji={3}}	0.1483	0.8974	2.6474
{RYG.Star={3},RYB.Emoji={3}} =>{RYB.Star={3}}	0.1483	0.8974	3.9961
{RYB.Emoji={3},RYB.Star={3}} =>{RYG.Star={3}}	0.1483	0.8333	3.7821
{RYG.Star={3},RYG.Emoji={3}} =>{RYB.Emoji={3}}	0.1525	0.9	2.655
{RYG.Star={3},RYB.Emoji={3}} =>{RYG.Emoji={3}}	0.1525	0.9231	2.8664
{RYG.Emoji={3},RYB.Star={3}} =>{RYB.Emoji={3}}	0.1653	0.8864	2.6148
{RYB.Emoji={3},RYB.Star={3}} =>{RYG.Emoji={3}}	0.1653	0.9286	2.8835
{Baseline={3},RYG.Emoji={3}} =>{RYB.Emoji={3}}	0.1441	0.8293	2.4463
{RYB.Emoji={3},RYB.Star={4,5}} =>{Baseline={3}}	0.1229	0.8788	2.8805
{RYG.Star={4,5},RYB.Emoji={3}} =>{Baseline={3}}	0.1102	0.8125	2.6632
{Baseline={3},RYB.Star={4,5}} =>{RYG.Star={4,5}}	0.1568	0.8409	1.4381
{Baseline={3},RYG.Star={4,5}} =>{RYB.Star={4,5}}	0.1568	0.9024	1.5546
{RYB.Emoji={3},RYB.Star={4,5}} =>{RYG.Star={4,5}}	0.1186	0.8485	1.451
{RYG.Star={4,5},RYB.Emoji={3}} =>{RYB.Star={4,5}}	0.1186	0.875	1.5073
{Baseline={4,5},RYB.Emoji={4,5}} =>{RYG.Emoji={4,5}}	0.3093	0.9359	1.9721
{RYG.Emoji={4,5},RYB.Emoji={4,5}} =>{Baseline={4,5}}	0.3093	0.8111	1.8767
{Baseline={4,5},RYG.Emoji={4,5}} =>{RYB.Emoji={4,5}}	0.3093	0.9605	2.2897
{Baseline={4,5},RYB.Emoji={4,5}} =>{RYB.Star={4,5}}	0.3093	0.9359	1.6122
{RYB.Emoji={4,5},RYB.Star={4,5}} =>{Baseline={4,5}}	0.3093	0.8022	1.8561
{Baseline={4,5},RYB.Star={4,5}} =>{RYB.Emoji={4,5}}	0.3093	0.9359	2.231
{Baseline={4,5},RYB.Emoji={4,5}} =>{RYG.Star={4,5}}	0.3136	0.9487	1.6224
{RYG.Star={4,5},RYB.Emoji={4,5}} =>{Baseline={4,5}}	0.3136	0.8132	1.8815
{Baseline={4,5},RYG.Star={4,5}} =>{RYB.Emoji={4,5}}	0.3136	0.9136	2.1778
{RYG.Emoji={4,5},RYB.Emoji={4,5}} =>{RYB.Star={4,5}}	0.3686	0.9667	1.6652
{RYB.Emoji={4,5},RYB.Star={4,5}} =>{RYG.Emoji={4,5}}	0.3686	0.956	2.0145
{RYG.Emoji={4,5},RYB.Star={4,5}} =>{RYB.Emoji={4,5}}	0.3686	0.8447	2.0135
{RYG.Emoji={4,5},RYB.Emoji={4,5}} =>{RYG.Star={4,5}}	0.3644	0.9556	1.6341
{RYG.Star={4,5},RYB.Emoji={4,5}} =>{RYG.Emoji={4,5}}	0.3644	0.9451	1.9914
{RYG.Star={4,5},RYG.Emoji={4,5}} =>{RYB.Emoji={4,5}}	0.3644	0.8515	2.0298
{RYB.Emoji={4,5},RYB.Star={4,5}} =>{RYG.Star={4,5}}	0.3686	0.956	1.635

Table C.2: Association rules for extroverts

Rules	Support	Confidence	Lift
{RYG.Star={4,5},RYB.Emoji={4,5}} =>{RYB.Star={4,5}}	0.3686	0.956	1.6469
{Baseline={4,5},RYG.Emoji={4,5}} =>{RYB.Star={4,5}}	0.3136	0.9737	1.6773
{Baseline={4,5},RYB.Star={4,5}} =>{RYG.Emoji={4,5}}	0.3136	0.9487	1.9991
{Baseline={4,5},RYG.Emoji={4,5}} =>{RYG.Star={4,5}}	0.3093	0.9605	1.6426
{Baseline={4,5},RYG.Star={4,5}} =>{RYG.Emoji={4,5}}	0.3093	0.9012	1.899
{Baseline={4,5},RYB.Star={4,5}} =>{RYG.Star={4,5}}	0.322	0.9744	1.6663
{Baseline={4,5},RYG.Star={4,5}} =>{RYB.Star={4,5}}	0.322	0.9383	1.6163
{RYG.Emoji={4,5},RYB.Star={4,5}} =>{RYG.Star={4,5}}	0.411	0.9417	1.6105
{RYG.Star={4,5},RYG.Emoji={4,5}} =>{RYB.Star={4,5}}	0.411	0.9604	1.6544
{RYG.Star={4,5},RYB.Star={4,5}} =>{RYG.Emoji={4,5}}	0.411	0.776	1.6351
{RYG.Star={1,2},RYG.Emoji={1,2},RYB.Star={1,2}} => {RYB.Emoji={1,2}}	0.1271	0.9091	3.764
{RYG.Star={1,2},RYB.Emoji={1,2},RYB.Star={1,2}} => {RYG.Emoji={1,2}}	0.1271	0.9677	4.7581
{RYG.Star={1,2},RYG.Emoji={1,2},RYB.Emoji={1,2}} => {RYB.Star={1,2}}	0.1271	0.9375	4.8098
{RYG.Emoji={1,2},RYB.Emoji={1,2},RYB.Star={1,2}} => {RYG.Star={1,2}}	0.1271	0.9091	4.664
{Baseline={1,2},RYG.Star={1,2},RYB.Star={1,2}} => {RYG.Emoji={1,2}}	0.1017	0.8889	4.3704
{Baseline={1,2},RYG.Star={1,2},RYG.Emoji={1,2}} => {RYB.Star={1,2}}	0.1017	0.8889	4.5604
{Baseline={1,2},RYG.Emoji={1,2},RYB.Star={1,2}} => {RYG.Star={1,2}}	0.1017	0.8889	4.5604
{RYG.Star={1,2},RYG.Emoji={1,2},RYB.Emoji={1,2}} => {Baseline={1,2}}	0.1017	0.75	2.8548
{Baseline={1,2},RYG.Star={1,2},RYG.Emoji={1,2}} => {RYB.Emoji={1,2}}	0.1017	0.8889	3.6803
{Baseline={1,2},RYG.Star={1,2},RYB.Emoji={1,2}} => {RYG.Emoji={1,2}}	0.1017	0.9231	4.5385
{Baseline={1,2},RYG.Emoji={1,2},RYB.Emoji={1,2}} => {RYG.Star={1,2}}	0.1017	0.8	4.1043
{Baseline={1,2},RYG.Emoji={1,2},RYB.Star={1,2}} => {RYB.Emoji={1,2}}	0.1017	0.8889	3.6803
{Baseline={1,2},RYB.Emoji={1,2},RYB.Star={1,2}} => {RYG.Emoji={1,2}}	0.1017	0.8276	4.069
{Baseline={1,2},RYG.Emoji={1,2},RYB.Emoji={1,2}} => {RYB.Star={1,2}}	0.1017	0.8	4.1043
{RYG.Star={3},RYG.Emoji={3},RYB.Star={3}} => {RYB.Emoji={3}}	0.1441	0.9444	2.7861
{RYG.Star={3},RYB.Emoji={3},RYB.Star={3}} => {RYG.Emoji={3}}	0.1441	0.9714	3.0165
{RYG.Star={3},RYG.Emoji={3},RYB.Emoji={3}} => {RYB.Star={3}}	0.1441	0.9444	4.2055
{RYG.Emoji={3},RYB.Emoji={3},RYB.Star={3}} => {RYG.Star={3}}	0.1441	0.8718	3.9566
{Baseline={3},RYB.Emoji={3},RYB.Star={4,5}} => {RYG.Star={4,5}}	0.1017	0.8276	1.4153
{Baseline={3},RYG.Star={4,5},RYB.Emoji={3}} => {RYB.Star={4,5}}	0.1017	0.9231	1.5901

Table C.3: Association rules for extroverts

Rules	Support	Confidence	Lift
{RYG.Star={4,5},RYB.Emoji={3},RYB.Star={4,5}} => {Baseline={3}}	0.1017	0.8571	2.8095
{Baseline={4,5},RYG.Emoji={4,5},RYB.Emoji={4,5}} => {RYB.Star={4,5}}	0.3051	0.9863	1.699
{Baseline={4,5},RYB.Emoji={4,5},RYB.Star={4,5}} => {RYG.Emoji={4,5}}	0.3051	0.9863	2.0783
{RYG.Emoji={4,5},RYB.Emoji={4,5},RYB.Star={4,5}} => {Baseline={4,5}}	0.3051	0.8276	1.9148
{Baseline={4,5},RYG.Emoji={4,5},RYB.Star={4,5}} => {RYB.Emoji={4,5}}	0.3051	0.973	2.3194
{Baseline={4,5},RYG.Emoji={4,5},RYB.Emoji={4,5}} => {RYG.Star={4,5}}	0.3008	0.9726	1.6633
{Baseline={4,5},RYG.Star={4,5},RYB.Emoji={4,5}} => {RYG.Emoji={4,5}}	0.3008	0.9595	2.0217
{RYG.Star={4,5},RYG.Emoji={4,5},RYB.Emoji={4,5}} => {Baseline={4,5}}	0.3008	0.8256	1.9102
{Baseline={4,5},RYG.Star={4,5},RYG.Emoji={4,5}} => {RYB.Emoji={4,5}}	0.3008	0.9726	2.3185
{Baseline={4,5},RYB.Emoji={4,5},RYB.Star={4,5}} => {RYG.Star={4,5}}	0.3008	0.9726	1.6633
{Baseline={4,5},RYG.Star={4,5},RYB.Emoji={4,5}} => {RYB.Star={4,5}}	0.3008	0.9595	1.6528
{RYG.Star={4,5},RYB.Emoji={4,5},RYB.Star={4,5}} => {Baseline={4,5}}	0.3008	0.8161	1.8882
{Baseline={4,5},RYG.Star={4,5},RYB.Star={4,5}} => {RYB.Emoji={4,5}}	0.3008	0.9342	2.227
{RYG.Emoji={4,5},RYB.Emoji={4,5},RYB.Star={4,5}} => {RYG.Star={4,5}}	0.3559	0.9655	1.6512
{RYG.Star={4,5},RYG.Emoji={4,5},RYB.Emoji={4,5}} => {RYB.Star={4,5}}	0.3559	0.9767	1.6826
{RYG.Star={4,5},RYB.Emoji={4,5},RYB.Star={4,5}} => {RYG.Emoji={4,5}}	0.3559	0.9655	2.0345
{RYG.Star={4,5},RYG.Emoji={4,5},RYB.Star={4,5}} => {RYB.Emoji={4,5}}	0.3559	0.866	2.0644
{Baseline={4,5},RYG.Emoji={4,5},RYB.Star={4,5}} => {RYG.Star={4,5}}	0.3051	0.973	1.6639
{Baseline={4,5},RYG.Star={4,5},RYG.Emoji={4,5}} => {RYB.Star={4,5}}	0.3051	0.9863	1.699
{Baseline={4,5},RYG.Star={4,5},RYB.Star={4,5}} => {RYG.Emoji={4,5}}	0.3051	0.9474	1.9962
{Baseline={4,5},RYG.Emoji={4,5},RYB.Emoji={4,5}, RYB.Star={4,5}} =>{RYG.Star={4,5}}	0.2966	0.9722	1.6626
{Baseline={4,5},RYG.Star={4,5},RYG.Emoji={4,5}, RYB.Emoji={4,5}} =>{RYB.Star={4,5}}	0.2966	0.9859	1.6984
{Baseline={4,5},RYG.Star={4,5},RYB.Emoji={4,5}, RYB.Star={4,5}} =>{RYG.Emoji={4,5}}	0.2966	0.9859	2.0775
{RYG.Star={4,5},RYG.Emoji={4,5},RYB.Emoji={4,5}, RYB.Star={4,5}} =>{Baseline={4,5}}	0.2966	0.8333	1.9281
{Baseline={4,5},RYG.Star={4,5},RYG.Emoji={4,5}, RYB.Star={4,5}} =>{RYB.Emoji={4,5}}	0.2966	0.9722	2.3176

Table C.4: Association rules for extroverts

C.2 214 Association Rules For Collectivists

Rules	Support	Confidence	Lift
{RYB.Star={1,2}} =>{RYG.Star={1,2}}	0.1573	0.8877	4.8027
{RYG.Star={1,2}} =>{RYB.Star={1,2}}	0.1573	0.8513	4.8027
{RYB.Star={1,2}} =>{RYG.Emoji={1,2}}	0.1555	0.877	4.7206
{RYG.Emoji={1,2}} =>{RYB.Star={1,2}}	0.1555	0.8367	4.7206
{RYB.Star={1,2}} =>{RYB.Emoji={1,2}}	0.1611	0.9091	4.589
{RYB.Emoji={1,2}} =>{RYB.Star={1,2}}	0.1611	0.8134	4.589
{RYB.Star={1,2}} =>{Baseline={1,2}}	0.1393	0.7861	3.6859
{RYG.Star={1,2}} =>{RYG.Emoji={1,2}}	0.1583	0.8564	4.6098
{RYG.Emoji={1,2}} =>{RYG.Star={1,2}}	0.1583	0.852	4.6098
{RYG.Star={1,2}} =>{RYB.Emoji={1,2}}	0.1583	0.8564	4.323
{RYB.Emoji={1,2}} =>{RYG.Star={1,2}}	0.1583	0.799	4.323
{RYG.Star={1,2}} =>{Baseline={1,2}}	0.1441	0.7795	3.6549
{RYG.Emoji={1,2}} =>{RYB.Emoji={1,2}}	0.163	0.8776	4.4297
{RYB.Emoji={1,2}} =>{RYG.Emoji={1,2}}	0.163	0.823	4.4297
{RYG.Emoji={1,2}} =>{Baseline={1,2}}	0.1403	0.7551	3.5406
{RYG.Star={3}} =>{RYB.Emoji={3}}	0.1858	0.7778	3.1682
{RYB.Emoji={3}} =>{RYG.Star={3}}	0.1858	0.7568	3.1682
{RYG.Star={3}} =>{RYB.Star={3}}	0.2	0.8373	3.3716
{RYB.Star={3}} =>{RYG.Star={3}}	0.2	0.8053	3.3716
{RYG.Star={3}} =>{RYG.Emoji={3}}	0.1962	0.8214	3.0731
{RYB.Emoji={3}} =>{RYB.Star={3}}	0.1915	0.7799	3.1405
{RYB.Star={3}} =>{RYB.Emoji={3}}	0.1915	0.771	3.1405
{RYB.Emoji={3}} =>{RYG.Emoji={3}}	0.2009	0.8185	3.0622
{RYG.Emoji={3}} =>{RYB.Emoji={3}}	0.2009	0.7518	3.0622
{RYB.Star={3}} =>{RYG.Emoji={3}}	0.2047	0.8244	3.0843
{RYG.Emoji={3}} =>{RYB.Star={3}}	0.2047	0.766	3.0843
{Baseline={4,5}} =>{RYG.Emoji={4,5}}	0.4521	0.8442	1.5436
{RYG.Emoji={4,5}} =>{Baseline={4,5}}	0.4521	0.8267	1.5436
{Baseline={4,5}} =>{RYB.Emoji={4,5}}	0.4521	0.8442	1.5173
{RYB.Emoji={4,5}} =>{Baseline={4,5}}	0.4521	0.8126	1.5173
{Baseline={4,5}} =>{RYB.Star={4,5}}	0.4588	0.8566	1.4913
{RYB.Star={4,5}} =>{Baseline={4,5}}	0.4588	0.7987	1.4913
{Baseline={4,5}} =>{RYG.Star={4,5}}	0.4607	0.8602	1.4926
{RYG.Star={4,5}} =>{Baseline={4,5}}	0.4607	0.7993	1.4926
{RYG.Emoji={4,5}} =>{RYB.Emoji={4,5}}	0.509	0.9307	1.6727
{RYB.Emoji={4,5}} =>{RYG.Emoji={4,5}}	0.509	0.9148	1.6727
{RYG.Emoji={4,5}} =>{RYB.Star={4,5}}	0.5185	0.948	1.6504
{RYB.Star={4,5}} =>{RYG.Emoji={4,5}}	0.5185	0.9026	1.6504
{RYG.Emoji={4,5}} =>{RYG.Star={4,5}}	0.5147	0.9411	1.633
{RYG.Star={4,5}} =>{RYG.Emoji={4,5}}	0.5147	0.8931	1.633
{RYB.Emoji={4,5}} =>{RYB.Star={4,5}}	0.5251	0.9438	1.6431
{RYB.Star={4,5}} =>{RYB.Emoji={4,5}}	0.5251	0.9142	1.6431
{RYB.Emoji={4,5}} =>{RYG.Star={4,5}}	0.5261	0.9455	1.6406
{RYG.Star={4,5}} =>{RYB.Emoji={4,5}}	0.5261	0.9128	1.6406
{RYB.Star={4,5}} =>{RYG.Star={4,5}}	0.5412	0.9422	1.635
{RYG.Star={4,5}} =>{RYB.Star={4,5}}	0.5412	0.9391	1.635

Table C.5: Association rules for collectivists

Rules	Support	Confidence	Lift
{RYG.Star={1,2},RYB.Star={1,2}} =>{RYG.Emoji={1,2}}	0.1469	0.9337	5.026
{RYG.Emoji={1,2},RYB.Star={1,2}} =>{RYG.Star={1,2}}	0.1469	0.9451	5.1134
{RYG.Star={1,2},RYG.Emoji={1,2}} =>{RYB.Star={1,2}}	0.1469	0.9281	5.2363
{RYG.Star={1,2},RYB.Star={1,2}} =>{RYB.Emoji={1,2}}	0.1469	0.9337	4.7134
{RYB.Emoji={1,2},RYB.Star={1,2}} =>{RYG.Star={1,2}}	0.1469	0.9118	4.9329
{RYG.Star={1,2},RYB.Emoji={1,2}} =>{RYB.Star={1,2}}	0.1469	0.9281	5.2363
{RYG.Star={1,2},RYB.Star={1,2}} =>{Baseline={1,2}}	0.128	0.8133	3.8133
{Baseline={1,2},RYB.Star={1,2}} =>{RYG.Star={1,2}}	0.128	0.9184	4.9686
{Baseline={1,2},RYG.Star={1,2}} =>{RYB.Star={1,2}}	0.128	0.8882	5.0107
{RYG.Emoji={1,2},RYB.Star={1,2}} =>{RYB.Emoji={1,2}}	0.1469	0.9451	4.7708
{RYB.Emoji={1,2},RYB.Star={1,2}} =>{RYG.Emoji={1,2}}	0.1469	0.9118	4.9077
{RYG.Emoji={1,2},RYB.Emoji={1,2}} =>{RYB.Star={1,2}}	0.1469	0.9012	5.0841
{RYG.Emoji={1,2},RYB.Star={1,2}} =>{Baseline={1,2}}	0.1251	0.8049	3.774
{Baseline={1,2},RYB.Star={1,2}} =>{RYG.Emoji={1,2}}	0.1251	0.898	4.8334
{Baseline={1,2},RYG.Emoji={1,2}} =>{RYB.Star={1,2}}	0.1251	0.8919	5.0318
{RYB.Emoji={1,2},RYB.Star={1,2}} =>{Baseline={1,2}}	0.127	0.7882	3.6959
{Baseline={1,2},RYB.Star={1,2}} =>{RYB.Emoji={1,2}}	0.127	0.9116	4.6014
{Baseline={1,2},RYB.Emoji={1,2}} =>{RYB.Star={1,2}}	0.127	0.8874	5.0066
{RYG.Star={1,2},RYG.Emoji={1,2}} =>{RYB.Emoji={1,2}}	0.1488	0.9401	4.7456
{RYG.Star={1,2},RYB.Emoji={1,2}} =>{RYG.Emoji={1,2}}	0.1488	0.9401	5.0603
{RYG.Emoji={1,2},RYB.Emoji={1,2}} =>{RYG.Star={1,2}}	0.1488	0.9128	4.9384
{RYG.Star={1,2},RYG.Emoji={1,2}} =>{Baseline={1,2}}	0.127	0.8024	3.7623
{Baseline={1,2},RYG.Star={1,2}} =>{RYG.Emoji={1,2}}	0.127	0.8816	4.7452
{Baseline={1,2},RYG.Emoji={1,2}} =>{RYG.Star={1,2}}	0.127	0.9054	4.8985
{RYG.Star={1,2},RYB.Emoji={1,2}} =>{Baseline={1,2}}	0.127	0.8024	3.7623
{Baseline={1,2},RYG.Star={1,2}} =>{RYB.Emoji={1,2}}	0.127	0.8816	4.4501
{Baseline={1,2},RYB.Emoji={1,2}} =>{RYG.Star={1,2}}	0.127	0.8874	4.8012
{RYG.Emoji={1,2},RYB.Emoji={1,2}} =>{Baseline={1,2}}	0.1261	0.7733	3.6257
{Baseline={1,2},RYG.Emoji={1,2}} =>{RYB.Emoji={1,2}}	0.1261	0.8986	4.5362
{Baseline={1,2},RYB.Emoji={1,2}} =>{RYG.Emoji={1,2}}	0.1261	0.8808	4.741
{RYG.Star={3},RYB.Emoji={3}} =>{RYB.Star={3}}	0.1697	0.9133	3.6775
{RYG.Star={3},RYB.Star={3}} =>{RYB.Emoji={3}}	0.1697	0.8483	3.4556
{RYB.Emoji={3},RYB.Star={3}} =>{RYG.Star={3}}	0.4607	0.8861	3.7098
{Baseline={3},RYG.Star={3}} =>{RYB.Emoji={3}}	0.509	0.8267	3.3673
{Baseline={3},RYB.Emoji={3}} =>{RYG.Star={3}}	0.509	0.8611	3.605
{RYG.Star={3},RYB.Emoji={3}} =>{RYG.Emoji={3}}	0.5185	0.9031	3.3785
{RYG.Star={3},RYG.Emoji={3}} =>{RYB.Emoji={3}}	0.5185	0.8551	3.483
{RYG.Emoji={3},RYB.Emoji={3}} =>{RYG.Star={3}}	0.5147	0.8349	3.4953
{Baseline={3},RYG.Star={3}} =>{RYB.Star={3}}	0.5251	0.8933	3.5972
{Baseline={3},RYB.Star={3}} =>{RYG.Star={3}}	0.5251	0.8933	3.7399
{RYG.Star={3},RYB.Star={3}} =>{RYG.Emoji={3}}	0.5261	0.9005	3.3688
{RYG.Star={3},RYG.Emoji={3}} =>{RYB.Star={3}}	0.5261	0.9179	3.696
{RYG.Emoji={3},RYB.Star={3}} =>{RYG.Star={3}}	0.5412	0.8796	3.6826
{Baseline={3},RYG.Star={3}} =>{RYG.Emoji={3}}	0.5412	0.8667	3.2423
{Baseline={3},RYG.Emoji={3}} =>{RYG.Star={3}}	0.1469	0.7975	3.3389
{Baseline={3},RYB.Emoji={3}} =>{RYB.Star={3}}	0.1469	0.8472	3.4115
{Baseline={3},RYB.Star={3}} =>{RYB.Emoji={3}}	0.1469	0.8133	3.313
{RYB.Emoji={3},RYB.Star={3}} =>{RYG.Emoji={3}}	0.1469	0.9208	3.4448
{RYG.Emoji={3},RYB.Emoji={3}} =>{RYB.Star={3}}	0.128	0.8774	3.5329
{RYG.Emoji={3},RYB.Star={3}} =>{RYB.Emoji={3}}	0.128	0.8611	3.5076

Table C.6: Association rules for collectivists

Rules	Support	Confidence	Lift
{Baseline={3},RYB.Emoji={3}} =>{RYG.Emoji={3}}	0.128	0.9167	3.4294
{Baseline={3},RYG.Emoji={3}} =>{RYB.Emoji={3}}	0.1469	0.8098	3.2987
{Baseline={3},RYB.Star={3}} =>{RYG.Emoji={3}}	0.1469	0.8733	3.2673
{Baseline={3},RYG.Emoji={3}} =>{RYB.Star={3}}	0.1251	0.8037	3.2362
{Baseline={3},RYG.Emoji={4,5}} =>{RYB.Emoji={4,5}}	0.1251	0.8219	1.4772
{Baseline={3},RYG.Emoji={4,5}} =>{RYB.Star={4,5}}	0.127	0.8082	1.407
{Baseline={3},RYG.Emoji={4,5}} =>{RYG.Star={4,5}}	0.1488	0.7945	1.3786
{Baseline={3},RYB.Emoji={4,5}} =>{RYB.Star={4,5}}	0.1488	0.7952	1.3843
{Baseline={3},RYB.Emoji={4,5}} =>{RYG.Star={4,5}}	0.127	0.8313	1.4425
{Baseline={3},RYG.Star={4,5}} =>{RYB.Emoji={4,5}}	0.127	0.7753	1.3934
{Baseline={3},RYB.Star={4,5}} =>{RYG.Star={4,5}}	0.127	0.8333	1.446
{Baseline={3},RYG.Star={4,5}} =>{RYB.Star={4,5}}	0.127	0.8427	1.4671
{Baseline={4,5},RYG.Emoji={4,5}} =>{RYB.Emoji={4,5}}	0.127	0.9602	1.7257
{Baseline={4,5},RYB.Emoji={4,5}} =>{RYG.Emoji={4,5}}	0.1261	0.9602	1.7556
{RYG.Emoji={4,5},RYB.Emoji={4,5}} =>{Baseline={4,5}}	0.1261	0.8529	1.5926
{Baseline={4,5},RYG.Emoji={4,5}} =>{RYB.Star={4,5}}	0.1261	0.9748	1.6971
{Baseline={4,5},RYB.Star={4,5}} =>{RYG.Emoji={4,5}}	0.1697	0.9607	1.7566
{RYG.Emoji={4,5},RYB.Star={4,5}} =>{Baseline={4,5}}	0.1697	0.8501	1.5873
{Baseline={4,5},RYG.Emoji={4,5}} =>{RYG.Star={4,5}}	0.4389	0.9706	1.6843
{Baseline={4,5},RYG.Star={4,5}} =>{RYG.Emoji={4,5}}	0.4389	0.9527	1.7419
{RYG.Star={4,5},RYG.Emoji={4,5}} =>{Baseline={4,5}}	0.4389	0.8527	1.5922
{Baseline={4,5},RYB.Emoji={4,5}} =>{RYB.Star={4,5}}	0.4398	0.9727	1.6935
{Baseline={4,5},RYB.Star={4,5}} =>{RYB.Emoji={4,5}}	0.4398	0.9587	1.723
{RYB.Emoji={4,5},RYB.Star={4,5}} =>{Baseline={4,5}}	0.4398	0.8375	1.5639
{Baseline={4,5},RYB.Emoji={4,5}} =>{RYG.Star={4,5}}	0.4389	0.9706	1.6843
{Baseline={4,5},RYG.Star={4,5}} =>{RYB.Emoji={4,5}}	0.4389	0.9527	1.7122
{RYG.Star={4,5},RYB.Emoji={4,5}} =>{Baseline={4,5}}	0.4389	0.8342	1.5577
{Baseline={4,5},RYB.Star={4,5}} =>{RYG.Star={4,5}}	0.4445	0.969	1.6814
{Baseline={4,5},RYG.Star={4,5}} =>{RYB.Star={4,5}}	0.4445	0.965	1.68
{RYG.Star={4,5},RYB.Star={4,5}} =>{Baseline={4,5}}	0.4445	0.8214	1.5337
{RYG.Emoji={4,5},RYB.Emoji={4,5}} =>{RYB.Star={4,5}}	0.4957	0.9739	1.6955
{RYG.Emoji={4,5},RYB.Star={4,5}} =>{RYB.Emoji={4,5}}	0.4957	0.9561	1.7184
{RYB.Emoji={4,5},RYB.Star={4,5}} =>{RYG.Emoji={4,5}}	0.4957	0.944	1.7261
{RYG.Emoji={4,5},RYB.Emoji={4,5}} =>{RYG.Star={4,5}}	0.4967	0.9758	1.6932
{RYG.Star={4,5},RYG.Emoji={4,5}} =>{RYB.Emoji={4,5}}	0.4967	0.965	1.7344
{RYG.Star={4,5},RYB.Emoji={4,5}} =>{RYG.Emoji={4,5}}	0.4967	0.9441	1.7263
{RYG.Emoji={4,5},RYB.Star={4,5}} =>{RYG.Star={4,5}}	0.5005	0.9653	1.6749
{RYG.Star={4,5},RYG.Emoji={4,5}} =>{RYB.Star={4,5}}	0.5005	0.9724	1.6928
{RYG.Star={4,5},RYB.Star={4,5}} =>{RYG.Emoji={4,5}}	0.5005	0.9247	1.6907
{RYB.Emoji={4,5},RYB.Star={4,5}} =>{RYG.Star={4,5}}	0.5109	0.9729	1.6882
{RYG.Star={4,5},RYB.Emoji={4,5}} =>{RYB.Star={4,5}}	0.5109	0.9712	1.6907
{RYG.Star={4,5},RYB.Star={4,5}} =>{RYB.Emoji={4,5}}	0.5109	0.944	1.6966
{RYG.Star={1,2},RYG.Emoji={1,2},RYB.Star={1,2}} => {RYB.Emoji={1,2}}	0.1412	0.9613	4.8524
{RYG.Star={1,2},RYB.Emoji={1,2},RYB.Star={1,2}} => {RYG.Emoji={1,2}}	0.1412	0.9613	5.1743
{RYG.Emoji={1,2},RYB.Emoji={1,2},RYB.Star={1,2}} => {RYG.Star={1,2}}	0.1412	0.9613	5.2008
{RYG.Star={1,2},RYG.Emoji={1,2},RYB.Emoji={1,2}} => {RYB.Star={1,2}}	0.1412	0.949	5.3542

Table C.7: Association rules for collectivists

Rules	Support	Confidence	Lift
{RYG.Star={1,2},RYG.Emoji={1,2},RYB.Star={1,2}} => {Baseline={1,2}}	0.1194	0.8129	3.8116
{Baseline={1,2},RYG.Star={1,2},RYB.Star={1,2}} => {RYG.Emoji={1,2}}	0.1194	0.9333	5.0238
{Baseline={1,2},RYG.Emoji={1,2},RYB.Star={1,2}} => {RYG.Star={1,2}}	0.1194	0.9545	5.1643
{Baseline={1,2},RYG.Star={1,2},RYG.Emoji={1,2}} => {RYB.Star={1,2}}	0.1194	0.9403	5.3049
{RYG.Star={1,2},RYB.Emoji={1,2},RYB.Star={1,2}} => {Baseline={1,2}}	0.1194	0.8129	3.8116
{Baseline={1,2},RYG.Star={1,2},RYB.Star={1,2}} => {RYB.Emoji={1,2}}	0.1194	0.9333	4.7113
{Baseline={1,2},RYB.Emoji={1,2},RYB.Star={1,2}} => {RYG.Star={1,2}}	0.1194	0.9403	5.0873
{Baseline={1,2},RYG.Star={1,2},RYB.Emoji={1,2}} => {RYB.Star={1,2}}	0.1194	0.9403	5.3049
{RYG.Emoji={1,2},RYB.Emoji={1,2},RYB.Star={1,2}} => {Baseline={1,2}}	0.1185	0.8065	3.7814
{Baseline={1,2},RYG.Emoji={1,2},RYB.Star={1,2}} => {RYB.Emoji={1,2}}	0.1185	0.947	4.7802
{Baseline={1,2},RYB.Emoji={1,2},RYB.Star={1,2}} => {RYG.Emoji={1,2}}	0.1185	0.9328	5.0211
{Baseline={1,2},RYG.Emoji={1,2},RYB.Emoji={1,2}} => {RYB.Star={1,2}}	0.1185	0.9398	5.3024
{RYG.Star={1,2},RYG.Emoji={1,2},RYB.Emoji={1,2}} => {Baseline={1,2}}	0.1204	0.8089	3.7929
{Baseline={1,2},RYG.Star={1,2},RYG.Emoji={1,2}} => {RYB.Emoji={1,2}}	0.1204	0.9478	4.7842
{Baseline={1,2},RYG.Star={1,2},RYB.Emoji={1,2}} => {RYG.Emoji={1,2}}	0.1204	0.9478	5.1015
{Baseline={1,2},RYG.Emoji={1,2},RYB.Emoji={1,2}} => {RYG.Star={1,2}}	0.1204	0.9549	5.1662
{Baseline={3},RYG.Star={3},RYB.Emoji={3}} => {RYB.Star={3}}	0.1081	0.9194	3.702
{Baseline={3},RYG.Star={3},RYB.Star={3}} => {RYB.Emoji={3}}	0.1081	0.8507	3.4654
{Baseline={3},RYB.Emoji={3},RYB.Star={3}} => {RYG.Star={3}}	0.1081	0.9344	3.912
{RYG.Star={3},RYB.Emoji={3},RYB.Star={3}} => {RYG.Emoji={3}}	0.1611	0.9497	3.553
{RYG.Star={3},RYG.Emoji={3},RYB.Emoji={3}} => {RYB.Star={3}}	0.1611	0.9605	3.8675
{RYG.Star={3},RYG.Emoji={3},RYB.Star={3}} => {RYB.Emoji={3}}	0.1611	0.8947	3.6446
{RYG.Emoji={3},RYB.Emoji={3},RYB.Star={3}} => {RYG.Star={3}}	0.1611	0.914	3.8264
{Baseline={3},RYG.Star={3},RYB.Emoji={3}} => {RYG.Emoji={3}}	0.11	0.9355	3.4998
{Baseline={3},RYG.Star={3},RYG.Emoji={3}} => {RYB.Emoji={3}}	0.11	0.8923	3.6347

Table C.8: Association rules for collectivists

Rules	Support	Confidence	Lift
{Baseline={3},RYG.Emoji={3},RYB.Emoji={3}} => {RYG.Star={3}}	0.11	0.8788	3.6791
{Baseline={3},RYG.Star={3},RYB.Star={3}} => {RYG.Emoji={3}}	0.1166	0.9179	3.434
{Baseline={3},RYG.Star={3},RYG.Emoji={3}} => {RYB.Star={3}}	0.1166	0.9462	3.8099
{Baseline={3},RYG.Emoji={3},RYB.Star={3}} => {RYG.Star={3}}	0.1166	0.9389	3.9308
{Baseline={3},RYB.Emoji={3},RYB.Star={3}} => {RYG.Emoji={3}}	0.1109	0.959	3.5878
{Baseline={3},RYG.Emoji={3},RYB.Emoji={3}} => {RYB.Star={3}}	0.1109	0.8864	3.5691
{Baseline={3},RYG.Emoji={3},RYB.Star={3}} => {RYB.Emoji={3}}	0.1109	0.8931	3.638
{Baseline={3},RYG.Emoji={4,5},RYB.Emoji={4,5}} => {RYG.Star={4,5}}	0.0521	0.9167	1.5906
{Baseline={3},RYG.Star={4,5},RYG.Emoji={4,5}} => {RYB.Emoji={4,5}}	0.0521	0.9483	1.7043
{Baseline={3},RYG.Star={4,5},RYB.Emoji={4,5}} => {RYG.Emoji={4,5}}	0.0521	0.7971	1.4574
{Baseline={3},RYB.Emoji={4,5},RYB.Star={4,5}} => {RYG.Star={4,5}}	0.0597	0.9545	1.6563
{Baseline={3},RYG.Star={4,5},RYB.Emoji={4,5}} => {RYB.Star={4,5}}	0.0597	0.913	1.5895
{Baseline={3},RYG.Star={4,5},RYB.Star={4,5}} => {RYB.Emoji={4,5}}	0.0597	0.84	1.5097
{Baseline={4,5},RYG.Emoji={4,5},RYB.Emoji={4,5}} => {RYB.Star={4,5}}	0.4294	0.9891	1.7219
{Baseline={4,5},RYG.Emoji={4,5},RYB.Star={4,5}} => {RYB.Emoji={4,5}}	0.4294	0.9742	1.7509
{Baseline={4,5},RYB.Emoji={4,5},RYB.Star={4,5}} => {RYG.Emoji={4,5}}	0.4294	0.9763	1.7851
{RYG.Emoji={4,5},RYB.Emoji={4,5},RYB.Star={4,5}} => {Baseline={4,5}}	0.4294	0.8662	1.6173
{Baseline={4,5},RYG.Emoji={4,5},RYB.Emoji={4,5}} => {RYG.Star={4,5}}	0.4265	0.9825	1.7049
{Baseline={4,5},RYG.Star={4,5},RYG.Emoji={4,5}} => {RYB.Emoji={4,5}}	0.4265	0.9719	1.7468
{Baseline={4,5},RYG.Star={4,5},RYB.Emoji={4,5}} => {RYG.Emoji={4,5}}	0.4265	0.9719	1.7771
{RYG.Star={4,5},RYG.Emoji={4,5},RYB.Emoji={4,5}} => {Baseline={4,5}}	0.4265	0.8588	1.6036
{Baseline={4,5},RYG.Emoji={4,5},RYB.Star={4,5}} => {RYG.Star={4,5}}	0.4313	0.9785	1.6979
{Baseline={4,5},RYG.Star={4,5},RYG.Emoji={4,5}} => {RYB.Star={4,5}}	0.4313	0.9827	1.7108
{Baseline={4,5},RYG.Star={4,5},RYB.Star={4,5}} => {RYG.Emoji={4,5}}	0.4313	0.9701	1.7738
{RYG.Star={4,5},RYG.Emoji={4,5},RYB.Star={4,5}} => {Baseline={4,5}}	0.4313	0.8617	1.6091

Table C.9: Association rules for collectivists

Rules	Support	Confidence	Lift
{Baseline={4,5},RYB.Emoji={4,5},RYB.Star={4,5}} => {RYG.Star={4,5}}	0.4303	0.9784	1.6978
{Baseline={4,5},RYG.Star={4,5},RYB.Emoji={4,5}} => {RYB.Star={4,5}}	0.4303	0.9806	1.7071
{Baseline={4,5},RYG.Star={4,5},RYB.Star={4,5}} => {RYB.Emoji={4,5}}	0.4303	0.968	1.7398
{RYG.Star={4,5},RYB.Emoji={4,5},RYB.Star={4,5}} => {Baseline={4,5}}	0.4303	0.959	3.5878
{RYG.Emoji={4,5},RYB.Emoji={4,5},RYB.Star={4,5}} => {RYG.Star={4,5}}	0.4872	0.8423	1.5728
{RYG.Star={4,5},RYG.Emoji={4,5},RYB.Emoji={4,5}} => {RYB.Star={4,5}}	0.4872	0.9828	1.7053
{RYG.Star={4,5},RYG.Emoji={4,5},RYB.Star={4,5}} => {RYB.Emoji={4,5}}	0.4872	0.9809	1.7077
{RYG.Star={4,5},RYB.Emoji={4,5},RYB.Star={4,5}} => {RYG.Emoji={4,5}}	0.4872	0.9735	1.7496
{RYG.Star={1,2},RYG.Emoji={1,2},RYB.Emoji={1,2}, RYB.Star={1,2}} =>{Baseline={1,2}}	0.1156	0.9536	1.7436
{Baseline={1,2},RYG.Star={1,2},RYG.Emoji={1,2}, RYB.Star={1,2}} =>{RYB.Emoji={1,2}}	0.1156	0.8188	3.8392
{Baseline={1,2},RYG.Star={1,2},RYB.Emoji={1,2}, RYB.Star={1,2}} =>{RYG.Emoji={1,2}}	0.1156	0.9683	4.8876
{Baseline={1,2},RYG.Emoji={1,2},RYB.Emoji={1,2}, RYB.Star={1,2}} =>{RYG.Star={1,2}}	0.1156	0.9683	5.2118
{Baseline={1,2},RYG.Star={1,2},RYG.Emoji={1,2}, RYB.Emoji={1,2}} =>{RYB.Star={1,2}}	0.1156	0.976	5.2804
{Baseline={3},RYG.Star={3},RYB.Emoji={3}, RYB.Star={3}} =>{RYG.Emoji={3}}	0.1052	0.9606	5.4196
{RYG.Star={3},RYG.Emoji={3},RYB.Emoji={3}, RYB.Star={3}} =>{Baseline={3}}	0.1052	0.9737	3.6427
{Baseline={3},RYG.Star={3},RYG.Emoji={3}, RYB.Star={3}} =>{RYB.Emoji={3}}	0.1052	0.9569	3.8532
{Baseline={3},RYG.Emoji={3},RYB.Emoji={3}, RYB.Star={3}} =>{RYG.Star={3}}	0.1052	0.9024	3.676
{Baseline={4,5},RYG.Emoji={4,5},RYB.Emoji={4,5}, RYB.Star={4,5}} =>{RYG.Star={4,5}}	0.4218	0.9487	3.9718
{Baseline={4,5},RYG.Star={4,5},RYG.Emoji={4,5}, RYB.Emoji={4,5}} =>{RYB.Star={4,5}}	0.4218	0.9823	1.7046
{Baseline={4,5},RYG.Star={4,5},RYG.Emoji={4,5}, RYB.Star={4,5}} =>{RYB.Emoji={4,5}}	0.4218	0.9889	1.7216
{Baseline={4,5},RYG.Star={4,5},RYB.Emoji={4,5}, RYB.Star={4,5}} =>{RYG.Emoji={4,5}}	0.4218	0.978	1.7578
{RYG.Star={4,5},RYG.Emoji={4,5},RYB.Emoji={4,5}, RYB.Star={4,5}} =>{Baseline={4,5}}	0.4218	0.9802	1.7922

Table C.10: Association rules for collectivists